Capacity Planning and Admission Control Policies for Intensive Care Units

by

Wongsakorn Chaiwanon

B.Eng., Chulalongkorn University (2007)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author .....
Sloan School of Management
July 31, 2010

Certified by .....
David Gamarnik
Associate Professor
Thesis Supervisor

Certified by .....
Retsef Levi
Associate Professor
Thesis Supervisor

Accepted by .....
Patr Jaillet
Edmund K. Turne ofessor
Codirector, Operations Resea Center

Capacity Planning and Admission Control Policies for Intensive Care Units

by

Wongsakorn Chaiwanon

Submitted to the Sloan School of Management
on July 31, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

Poor management of the patient flow in intensive care units (ICUs) causes service rejections and presents significant challenges from the standpoint of capacity planning and management in ICUs. This thesis reports on the development of a simulation framework to study admission control polices that aim to decrease the rejection rate in the ICU at Children's Hospital Boston (CHB), and to provide predictions for the future state of the ICU system. To understand the patient flow process, we extensively analyze the arrival and length of stay (LOS) data from the ICU census. The simulation model for the ICU is developed based on the results from this statistical analysis as well as the currently-practiced scheduling and admission policies of the ICU at CHB. The model is validated to provide accurate estimates for important performance metrics such as rejection rates in the ICU.

The simulation model is used to study the performance of many admission control policies. The policies of our interest exploit "caps" to control the number of scheduled patients who are allowed to enter the ICU on a single day. In particular, we consider two cap-based policies: the uniform cap policy (UCP), which is the existing policy in CHB, and the service-specific cap policy (SSCP), which is originally proposed in this thesis. While the UCP implements caps on the total census of surgical patients, the SSCP utilizes the service-oriented heterogeneity of surgical patients' LOS and enforces caps on separate groups of surgical patients based on their average LOS. We show that the UCP can reduce the rejection rate in the ICU at the expense of extra waiting time of scheduled patients. The SSCP is shown to further decrease the rejection rate while increasing the waiting time compared to the UCP. We also demonstrate that the performance of both policies depends on the level of system utilization. In order to validate our results theoretically, a discrete-time queueing model for the ICU is developed and verified to provide estimates for performance measures that are consistent with the results from simulation.

Finally, we introduce the notion of state-dependent prediction, which aims to identify the likelihood of the future state of the ICU conditional on the information of a current state. Several experiments are conducted by simulation to study the

impact of a current state on a state in the future. According to our results, current state information can be useful in predicting the state of the ICU in the near future, but its impact gradually diminishes as the time difference between the present and future grows. Our major finding is that the probability of unit saturation at a certain future time can be determined almost entirely by the number of current patients who will leave the ICU after that time, regardless of the total number of patients who are currently staying in the unit. These results imply the potential development of adaptive cap-based policies that dynamically adjust caps according to the outcomes of state-dependent predictions.

Thesis Supervisor: David Gamarnik
Title: Associate Professor

Thesis Supervisor: Retsef Levi
Title: Associate Professor

Acknowledgments

Many individuals have significantly contributed to the success of this thesis. First and foremost, I am grateful to my supervisors, Professors David Gamarnik and Retsef Levi, for their dedication, enthusiasm, and continuous support throughout my years at MIT. David and Retsef taught and showed me how to develop originality and think critically about research problems. They have carefully read the thesis and provided many excellent advices for improving the quality of this work. Without their guidance and attention, I would not have been able to complete this research.

This thesis is based on collaborative work with Dr. Michael McManus from the Division of Critical Care Medicine at Children's Hospital Boston. Mike shared perceptive insights that help shape my understanding of the critical care unit system and discussed many solution techniques from a healthcare policy standpoint. I truly appreciate his excellent supervision and valuable feedback over the course of this project.

I wish to express my appreciation to friends who have made MIT my wonderful home for the past three years. My transition from Bangkok to Cambridge was smooth and welcome thanks to the excellent hospitality from Tanachai and Kampol. I gratefully acknowledge the very kind assistance from Tanachai in helping me settle down and guiding me through many hard times during the initial stage of my stay at MIT. I wish to thank the Thai community at MIT for sharing with me many joyful and memorable moments. My special thank goes to Watcharapan for his helpful advice and intellectual guidance on my academics and research interests. In addition, I have greatly enjoyed the friendship with many fellows from the ORC, especially Allison, Phil, Wei, Claudio, and Christ. I am also thankful to the staff members of the ORC (Laura, Paulette, and Andrew) for their help on numerous occasions.

I thank my girlfriend, Por, for the unwavering support she has given to me over years. She always brings me great joy every single day with her love and caring, and we have had so many wonderful times and journeys together for the past seven years. Finally, I owe the deepest gratitude to my mother and sister for the true love, for the

encouragement, for the understanding, and for listening to me when I need them the most.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis studies the patient flow in the ICU at Children's Hospital Boston (CHB) and develops simulation framework that allows us to test various admission control policies and make inference about the future state of the ICU. In this chapter, we provide problem motivation, discuss methodology, contributions, and literature survey, and give the outline of the thesis.

1.1 Motivation

Intensive care units (ICUs) provide critical care for critically ill patients. Requiring highly specialized medical resources that include staff and equipment, the ICU is one of the most expensive units in a hospital. The cost associated with ICUs in the US accounts for 15%- 20% of US hospital costs, which represents 38% of total US health-care costs (Gruenberg et al. [16]). The demand for ICUs is also rising as evidenced by the overcrowding in many hospitals' ICUs. The study in Green [11], for example, indicates that in 2003 90% of ICUs in New York state have insufficient capacity to properly provide critical care to their patients. Congestion and poor management of the flow of patients in ICUs causes the cancelation of scheduled surgery, diversions of emergency cases to other hospitals, and premature discharges. These consequences lead not only to detrimental effects on patient safety and quality of care, but also losses of hospital revenue.

As one of the largest pediatric ICUs in the area, the ICU at CHB has always encountered with the problem of service rejections that take place during overcrowding hours. The hospital has addressed this problem through various approaches. One of these is to increase the number of ICU beds to match the increasing demand for critical care. In particular, the capacity of the ICU at CHB was raised from 17 to 29 beds over the period from 2003-2008. The hospital also built a separate ICU for caring non-surgical patients in 2008 to better serve the demand from this type of patients.

Another attempt by the CHB to improve the flow of patients within the ICU is the implementation of an admission control policy. Since 2003, the unit has enforced a limit (cap) on the number of surgical patients requiring post-operative ICU beds that can be scheduled to a single day. The policy aims to reduce variability in the daily demand from ICU scheduled surgical patients. This is indeed shown to be a major cause of poor management of patient flow in the unit, according to the study in McManus et al. [29], which is based on the data from the ICU at CHB. Nonetheless, there has been no effort to track the impacts of this cap-based policy on either the demand variability or admission rate in the ICU since its first use in 2003. One of the goals of this thesis is to fill this gap.

Other admission control policies can be potentially implemented to increase the ICU throughput and efficiency. For example, taking advantage of the heterogeneity in the occupancy times of patients, the ICU could devise a variation of cap-based policies to specifically control the admission of patients with long lengths of stay (LOS). According to the reports by Stricker et al. [33] and Ryckman et al. [32], patients of this type consume most of ICU resources although they represent only a minority in the pool of total admitted patients. An effort to limit the number of long-stay patients that can be admitted per day would allow better distribution of ICU resources and lead to smaller rejection rates. Moreover, the ICU might consider integrating the knowledge of its current state into the development of an admission control policy. This way, instead of using static caps, the unit could exploit current state information to predict the future state of the system and dynamically adjust

caps accordingly.

In this thesis, we will systematically study the current practices and potential alternatives of the admission control policies outlined above in the context of the ICU at CHB.

1.2 Methods

We use the method of simulation as the main tool in modeling the dynamics of the ICU at CHB in this thesis. Since it is free from any particular kind of assumptions on any specific characteristic of a system, simulation offers the versatility in detailed modeling of complex stochastic systems as well as the flexibility to study complicated "what-if" scenarios. In case of an ICU, we can use a simulation model to investigate the performance of many related admission and scheduling policies in the unit under many variations of inflow demand, LOS, and the unit capacity. A well-developed model would also allow us to understand the behavior of the system in future time given any initial current state of the ICU.

From a modeling perspective, nonetheless, developing detailed simulation for an ICU system can be particularly challenging for a number of reasons. In many cases, data might be limited or incomplete to model related system components or processes. For example, we encounter the situation where the we have no access to the record of all medical patients who arrived to the ICU and need to instead approximate their true arrival rates. Difficulty also arises when admission/scheduling policies or other rules in an actual ICU system are so involved that it is infeasible to capture these exactly by simulation. Therefore, often times the process of developing a simulation model involves certain simplifying assumptions and approximation techniques, and it is necessary for such a model to be validated in order to prove its correctness.

As a first step in building a simulation model for the ICU at CHB, we perform extensive data analysis of the ICU census to understand the patient flow in the real unit. Our statistical analysis in particular focuses on the arrival and LOS data. Then, a discrete-event simulation model for the ICU is developed, and it is calibrated to fit

the environment, practices, and flow of patients in the ICU based on the results from the data analysis. Some assumptions about arrival rates as well as admission and scheduling policies are made in the model. Finally, the simulation model is validated with the actual historical data from the ICU at CHB.

We next consider two different cap-based admission control policies. The first one, which we call the uniform cap policy (UCP), is the existing policy in the ICU at CHB. It uses caps to limit the total number of surgical cases requiring an ICU that can be scheduled per day in order to reduce the variability in surgical caseload. Then, based on the service-specific heterogeneity in the average LOS of ICU surgical patients, we propose the service-specific cap policy (SSCP) that uses caps to control the admission of scheduled surgical patients by service types, rather than the total number as currently implemented by the UCP. The goal of the SSCP is to limit the number of long-stay surgical patients that can be scheduled per day. Various performance measures including the rejection rates and mean waiting time are evaluated for both policies by our simulation model. We also study the trade-off between the rejection rates and the mean waiting time when the capacity of caps is adjusted, as well as the performance of both policies in the ICU with the varying degrees of system utilization. In addition, we investigate the decrease in the rejection rates when the LOS of each patient is reduced.

Then, we formulate a discrete-time queueing model as an analytical alternative to analyze the performance of the ICU at CHB. In particular, our model aims to capture the dynamics of the ICU system that uses caps in scheduling elective surgical patients. Certain simplifiying assumptions are made upon the arrivals and LOS to establish Markov property of the underlying stochastic processes. Then, the stationary probability of each state of the system can be solved numerically and is subsequently used to compute the rejection rate and the mean waiting time of the ICU system in the steady state. The results of these two performance measures are compared to those obtained from the simulation model.

Finally, we study the ICU state-dependent prediction problem that uses current state information to predict the probability distribution of a state in future time.

Given a current state of the ICU, we are particularly interested in the probability that the ICU will be fully-occupied in the future. The study of state-dependent prediction is divided into two parts based on the ICU's knowledge of current state information. In the first part, we assume that the departure times of current patients can be estimated only from the empirical distributions of the LOS conditional on their current LOS. Our goal is to investigate the impact of current state information on a future state as the time difference between the present and future is increased. In the second part, we assume that the every current patient can be assessed by the ICU as to whether he/she is leaving the unit before a certain future point in time. This assumption is made based on the fact that medical professionals can often come up with fairly accurate guesses about the remaining LOS of current patients after a certain period of monitoring their stays in the ICU. Various scenarios are tested by simulation to relate this additional knowledge of the departure times of current patients to the future state of the system.

The key questions that will be addressed in this research are as follows.

1. Is the implementation of the UCP able to reduce variability in scheduled surgery demand for intensive care in any significant way? Can the policy decrease the rejection rate in the ICU? What are the trade-offs involved?

2. Can the SSCP further decrease the demand variability as well as the rejection rate in the ICU when compared to the UCP?

3. Is current state information useful in predicting the future state of the ICU?

4. What is the impact of the departure times of current patients on the likelihood that the ICU will be full in the future?

1.3 Research Results

Our results are summarized as follows.

- We build a discrete-event simulation model and validate it to be an accurate model for the ICU at CHB.

- We show that the implementation of the UCP in the ICU reduces variability in scheduled surgery demand and decreases the rejection rate, though at the expense of longer mean waiting time of scheduled patients in the ICU.

- We show that the SSCP decreases both the demand variability and the rejection rate in the ICU further when compared with the UCP. Meanwhile, the SSCP results in longer mean waiting time of scheduled patients.

- We show that both the UCP and SSCP contribute the most to decreasing the rejection rate when the ICU is operating at approximately $70\% - 75\%$ utilization level.

- We show that decreasing the LOS of ICU patients even by a few hours on average can lead to the noticeable decrease in the rejection rate.

- We develop a discrete-time queueing model for the ICU at CHB. It is used to analyze the ICU system that implements the UCP. We show that the performance measures of the ICU computed from the queueing model at various cap levels are consistent with those from simulation. Thus, our model provides a viable alternative to the simulation approach.

- We introduce the framework of state-dependent prediction, which uses current state information to forecast the future state of the ICU. We demonstrate that a current state can indeed affect the likelihood of various states in the nearby future, but its impact on the state in the farther future gradually disappears. In addition, when the knowledge of current patients' remaining LOS is assumed, we show that the probability that the ICU will be full at a future point in time is virtually independent of the number of patients who are known to leave the unit before that time (short-stay patients). Instead, it depends almost entirely on the number of current patients who are going to stay in the unit through that time (long-stay patients).

1.4 Literature Review

Operations Research methods have been widely used to analyze and aid the decision-making in healthcare systems. Queueing theory and computer simulation are among the most popular Operations Research modeling techniques in healthcare when the behavior of the system considered is highly stochastic. The ICU is one of hospital units that are characterized by multiple sources of randomness, and its patient flow process is often modeled by queueing analysis and simulation.

Queueing theory methods have been central in several recent healthcare research papers. Chaussalet et al. [5], Jiang and Giachetti [17], and Koizumi et al. [21] analyze patient flow across the hospital using queueing networks. Asaduzzaman et al. [2] develop a loss network model that is used for capacity planning in the neonatal unit of a perinatal network in the United Kingdom. Subsequently, Asaduzzaman and Chaussalet [1] extend the model in Asaduzzaman et al. [2] by proposing a loss network queueing model that also captures the possibility of overflow in the same perinatal network. Tucker et al. [35] and Green et al. [13] consider the problem of staffing decisions in operating room (OR) and emergency department (ED), respectively, from a queueing perspective. In addition, Yankovic and Green [38] develop a queueing model to help identifying nurse staffing levels in hospital clinical units. Note that the works in Jiang and Giachetti [17], Koizumi et al. [21], and Tucker et al. [35] also construct simulation models to validate the results obtained from queueing analysis. One can refer to Green [12] for an overview of queueing theory methods for capacity management in hospitals.

Modeling ICU systems by queueing theory has been a subject of interest to many researchers. McManus et al. [30] demonstrate that an M/M/c/c queueing model accurately estimates the performance measures of the ICU at CHB compared to those from the actual data. Griffiths and Price-Lloyd [14] develop a multi-channel M/H/c/∞ model that assumes hyper-exponential service times in the ICU to capture the high variation in the LOS. Among recent studies, Kortbeek and van Dijk [24] use the results from M/G/c/c loss models to establish analytical bounds of rejection

probability in an Operating Theater-Intensive Care Unit (OT-ICU) tandem queue. Litvak et al. [26] propose an overflow model for cooperative capacity planning in a network of ICUs and use it to compute the number of required beds for any predefined acceptance rate. In addition, Dobson et al. [8] develop a queueing model for an ICU that, when overcrowded, diverts patients to other medical units to make room for new arrivals. Recently, Chan et al. [4] considered the discrete-time queueing dynamics of ICUs with patient readmission and developed an optimal discharge policy associated to the model.

Simulation has also been a primary approach in studying several healthcare units and centers. Lowery [27] develops a simulation model for a hospital and uses it to design a scheduling policy that reduces the variability in the daily census. Kolker [22] uses discrete-event simulation methodology to determine the rate of diversion in ED as a function of the upper limits of LOS. Based on the method of simulation, a series of authors Dexter et al. [7], Dexter and Traub [6], Tyler et al. [36], and van Houdenhoven et al. [37] investigate various approaches to scheduling elective surgery cases in order to increase the utilization in ORs. In addition, Ferreira et al. [9] adopt a framework of discrete-event simulation to analyze patient flow and identify strategies for improving the performance of a large surgical center in Rio de Janeiro, Brazil.

A substantial literature on healthcare systems focuses on developing simulation models for ICUs. Kim et al. [18] is among the first groups of authors that report on the development of a simulation model for the ICU of a public hospital in Hong Kong. The model is used to calculate performance measures and come up with capacity planning recommendations in that ICU. Subsequently in the work of Kim et al. [19], the same authors extend their previous simulation framework to study various bed-reservation policies for elective surgery cases in the same ICU. Simulation results show that the proposed policies can reduce the number of rejections, but at the same time lead to the increasing waiting time of scheduled patients. Ridge et al. [31] study bed capacity planning in Southampton General Hospital ICU and show via simulation a strong trade-off between the unit capacity and the number of rejections in the simulated ICU. Recently, Troy and Rosenberg [34] built a discrete-event simulation

model of the ICU at Jewish General Hospital, Montreal, and studied the correlation between the number of beds reserved for surgical patients and various performance measures in that ICU.

We now explore literature that is immediately to our work in this thesis. Ryckman et al. [32] study the framework of improving overall patient flow in the 35-bed pediatric ICU at Cincinnati Children's Hospital Medical Center by implementing a large number of new activities simultaneously in the ICU. Of particular interest are using caps for scheduling elective surgery cases to smoothen inflow demand and limiting the number of occupied beds per day for patients who are predicted to have long LOS. Note that the cap levels are adjusted according to the number of available ICU staff. Other methods include daily anticipation of demand on the next day, daily forecast about the number of discharged patients, the prediction of patient LOS, and the use of simulation to predict bed occupancy in the unit. While the result is promising in that cancelations and diversions have become uncommon, it is not clear to what extent caps or the effort in controlling bed allocation to long-stay patients contributed to this outcome, since a variety of strategies have been attempted at the same time. In addition, the authors do not provide the diversion/cancellation rates in the period prior to the implementation of the new polices, and it remains unanswered how significant the impact of these policies was to improving throughput of the ICU.

In Kolker [23], the author uses simulation to study the impacts of caps on the patient flow of an ICU. The goal of this research is to determine the size of caps that reduces the diversion rate in the ICU with fixed capacity to an acceptable level. The modeled system includes a 49-bed ICU with two extra beds left for emergency admissions, and the simulation is conducted over the period of 18 weeks. The simulation model is validated to give accurate estimates for diversion rates. The author shows that the variability in the inflow demand from scheduled surgeries can be reduced after enforcing caps. More importantly, he demonstrates a very significant improvement in the diversion percentage, from 10.5% to 1.5%, when the cap of four cases per day is used in the simulated ICU. To obtain this striking result, this paper considers a situation where all demand is assumed to be known in advance throughout the fu-

ture time horizon. Then, the cap-based policy is implemented in a way that surgical patients who were originally scheduled to a day on which the total load already exceeds the cap capacity will be rescheduled to other days on which the caseload is the lightest in order to smoothen the demand for an ICU as much as possible. Because of this policy, scheduled surgical patients might have to wait longer than two months away from their original waiting times to enter the ICU. To alleviate this potentially long additional waiting time, the author considers the exact same problem, except that now elective patients cannot be delayed by more than two weeks and the cap is raised to five cases per day. He shows that caps become less effective in this case, as the diversion rate only goes down from 10.5% to 8%. We need to emphasize that the studies in this research are conducted based on unrealistic scenarios, since in reality future demand cannot be completely determined beforehand. The corresponding results might provide an upper bound for the improvement in diversion rates gained from caps, yet they are unlikely to be the case in an actual ICU system.

1.5 Thesis Outline

This thesis is organized as follows. Chapter 2 describes the environment of the ICU at CHB and reports on the statistical analysis of arrivals and LOS from the ICU census. Chapter 3 discusses the development and validation of the simulation model for the ICU at CHB. Chapter 4 studies the cap-based admission control policies and evaluates their performance in the ICU by simulation. In Chapter 5, we formulate the queueing model for the ICU system and compare the performance measures obtained from the model with those from simulation. Chapter 6 studies the problem of state-dependent prediction. Finally, Chapter 7 provides the summary of the results in this thesis and directions for further research.

Chapter 2

ICU at Children's Hospital Boston

In this chapter, we describe the ICU environment at Children's Hospital Boston (CHB). In particular, we discuss the general detail of ICU patient characteristics, scheduling and admission policies, and significant changes that have been made in the unit throughout the years. In addition, we provide the statistical analysis on the arrivals and the length of stay (LOS) of ICU patients based on the data collected by the ICU personnel.

2.1 Populations

The ICU at CHB serves patients from a variety of sources. Most of the patients arrive from within the hospital itself, particularly the surgery and emergency departments. In addition, there are transferred patients from other hospitals that require critical care services. Patients who require care related to surgery issues are called *surgical* patients, while those who require care for other medical issues are called *medical* patients. All admissions to the ICU at CHB fall broadly into two categories according to the urgency of their requests.

1. **Scheduled patients**. This type of patients consists of surgical cases that do not require urgent surgery (elective surgery cases). They generally come from the surgery department and their surgery dates are scheduled by the booking office of the operating rooms (ORs).

2. **Emergency patients**. This type of patients consists of both surgical and medical cases that require urgent care and come to either the ORs or the ICU without any preplanned scheduling. Emergency patients come from the emergency department, floors, and transfers. The majority of emergency patients comprise medical patients.

2.2 Scheduling Policies

All elective patients must undergo a scheduling process to arrange dates and times for their surgery. They are generally, with some exceptions, scheduled to the most recent available block of surgeons according to the First-In-First-Out (FIFO) discipline. Block or block time is a preplanned fixed time in a given OR that is allocated to an individual surgeon. Via block time, a surgeon knows months in advance which day(s), times, and ORs (e.g., Monday, 8-17, OR#1) she will operate, and can schedule her cases accordingly.

Of all elective surgery patients, only some require critical care after surgery. Before 2003, there was no coordination between the ICU and the OR scheduling office, so elective surgical patients that require intensive care were scheduled arbitrarily. In fact, any number of cases could be booked on any day. The lack of control mechanism to schedule patients gave rise to particularly wide fluctuation in the daily demand from scheduled surgical caseload, which is shown to be an underlying cause for the limited access to the ICU (McManus et al. [29]).

This finding led to the implementation of an admission control policy by using *caps* (Kolker [23] and Ryckman et al. [32]) in the ICU at CHB. A cap is an administratively imposed limit on the number of daily scheduled surgeries that require ICU resources. It serves as a first-line attempt to smoothen the ICU demand generated by elective surgery cases. With caps being administrated, surgeons are not allowed to schedule and have to move cases if the number of surgical patients requiring post-surgery critical care that have been scheduled on a given day exceeds the cap.

We next describe the specifics of the cap-based policy that has been implemented

in the ICU at CHB since 2003. For each day of the week, there is a predefined cap to limit the *total* number of ICU elective surgeries that can be scheduled on that day. In this case, elective patients who require post-operative ICU beds are still scheduled by FIFO to the first available block of surgeons, except that the OR booking office is instructed to limit the *total* number of such cases that can be booked on that day according to the cap. We call this policy the **uniform cap policy** (UCP).

Motivated by the high consumption of ICU resources from patients with long LOS (Ryckman et al. [32] and Stricker et al. [33]), we propose the **service-specific cap policy** (SSCP) as an extension of the UCP. This policy is designed to allocate caps to limit the number of surgical patients from *separate groups* that can be scheduled on a single day, rather than the whole patients as currently executed in the UCP. Indeed, a cap of the SSCP aims to restrict the admission to the ICU of the long-stay types of surgical patients (the analysis of the LOS statistics by services is provided in Section 2.8.3). By enforcing the SSCP, elective surgery patients with a priori-known surgical services are scheduled by FIFO to the first available block time of surgeons as long as the number of scheduled cases on that day does not violate the cap restriction for the respective types of surgical services.

We will elaborate on both cap-based admission control policies and discuss their performance by means of simulation in Chapter 4.

2.3 Admission Policies

We next describe the admission policies and practices in the ICU at CHB. Of all the beds in the ICU, one bed is always reserved as a *crash bed*. The crash bed is used only in the situation where there is a newcoming patient who requires critical care and all other beds are already occupied. Once the crash bed is filled, the unit will immediately make a new crash bed available by diverting one of the patients, who can potentially leave the ICU by then, to another unit. As a result, the ICU with x beds with one crash bed is equivalent to the ICU with the capacity of $x - 1$ beds. The ICU is considered to be full if all other beds except the crash bed are occupied.

Scheduled surgery cases may be canceled or rescheduled prior to the surgery dates to accommodate surgeons' schedules or the need of patients themselves. Once the surgery starts, however, surgical patients that require critical care cannot be rejected and have to be admitted by the ICU. If a surgical patient arrives to a full ICU, chances are that the patient (or other surgical patients already in the ICU) will be diverted to other medical units within the hospital, such as the post anesthesia care unit (PACU) and the cardiac ICU, or that he will be admitted to the crash bed depending on the severity of his post-operative condition.

On the other hand, ICU medical patients can be rejected upon their requests for admissions. However, since medical patients are usually sicker than surgical ones and need an ICU immediately without prior warning, their diversions are not medically desirable. In case that a medical patient arrives to a full unit, the ICU always tries to make a bed available for her by diverting current surgical patients to other units. In fact, the unit is obligated to admit medical patients from within CHB, but can reject those who are transferred from referral hospitals if their illnesses can be cared in other units or hospitals.

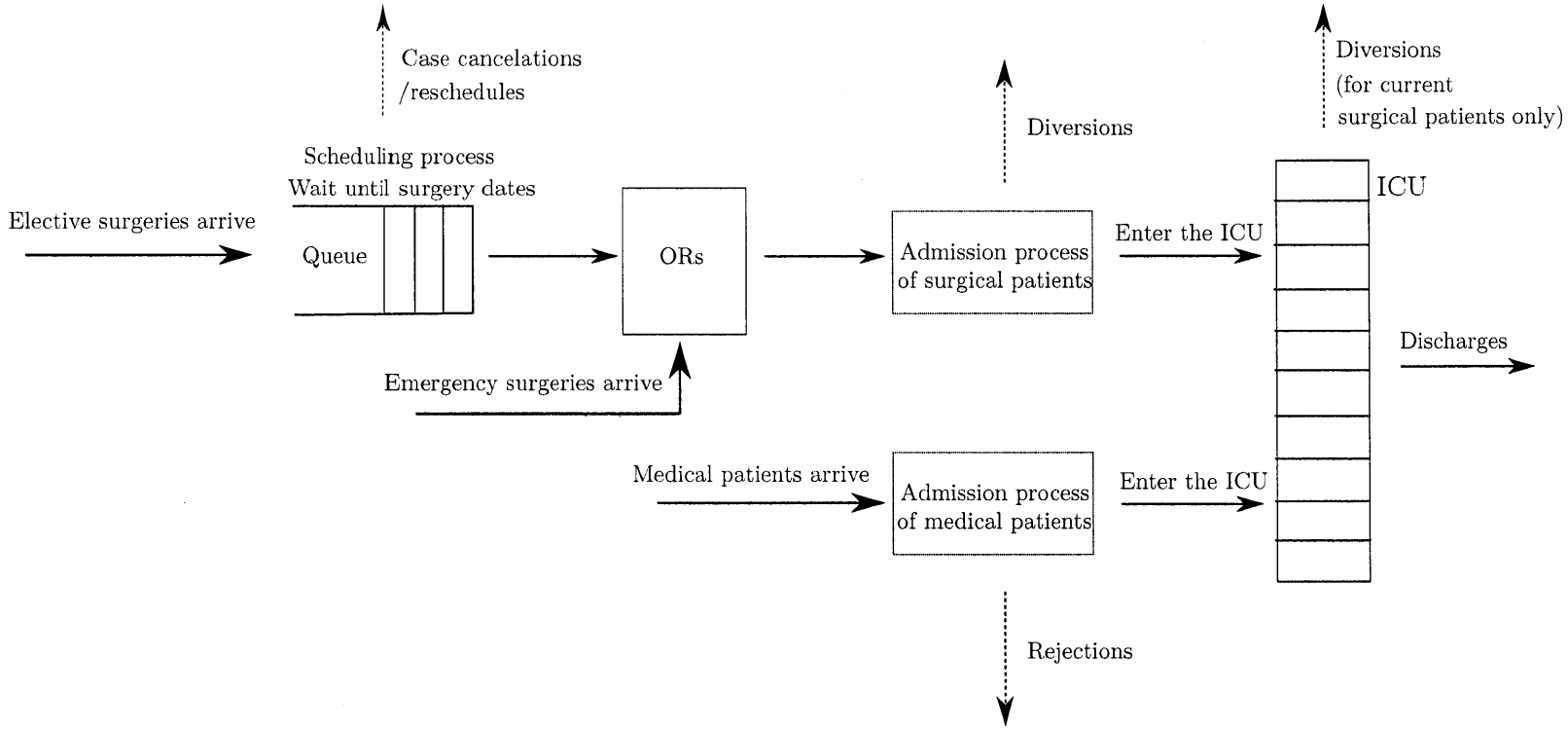


Figure 2-1: Patient flow in the ICU at CHB

2.4 Summary of Changes in Capacity and Scheduling Policies

The ICU at CHB has undergone many changes in the capacity and scheduling policies over the last decade to better serve the increasing demand for critical care. The important changes are summarized in Table 2.1

Date	Capacity	Scheduling policies
Before 2003	18 beds including one crash bed	No admission control policy
April 2003	-	Fully functional surgical cap at 5 cases per day
June 2005	23 beds including one crash bed	-
January 2007	29 beds including one crash bed	-
March 2008	29 beds plus a new medical ICU with 10 beds	Cap is increased to 6 cases per day.
June 2008	-	Cap is raised to 7 cases on Mondays and Tuesdays.

Table 2.1: The summary of the capacity and scheduling policies throughout years of the ICU at CHB

Regarding the new 10-bed medical ICU in 2008, there are no precise sets of rules to determine to which ICU a medical patient should be sent. The decision tends to be made on a case by case basis although most medical patients end up staying in the new medical ICU. Also, there is a preference to treat the more complex patients in the main ICU.

2.5 Data Analysis

We perform statistical analysis of the data from the ICU at CHB. The results provide us with an understanding of the arrivals and the length of stay (LOS) of the ICU patients. This is necessary for constructing a simulation model of the ICU. Our main findings are that the arrival rates and LOS vary by seasons of the year and that the

average LOS can be significantly different depending on surgical services.

2.5.1 Data

CHB has been collecting the data of their ICU patients since 1998. The data is available for patients who were admitted to the ICU and for surgical patients who were diverted to other units inside the hospital. However, it does not include medical patients who were rejected from the unit, except for 2000, in which year the hospital did track the number of medical rejections. For each ICU patient, the following records are used in our analysis:

- Admission dates and times
- The types of patients: surgical and medical. Surgical patients are also given their specific types of surgical services
- Discharge dates, times, and locations

The admission and departure dates and times are used to analyze the arrivals and LOS statistics of ICU patients. We classify patients based on their services rather than symptoms. This is because patients' types of services are perfectly determined upon their arrivals, while the symptom identification needs diagnosis, which might not always be the correct one at the time of admission and could lead to the wrong classification of patients.

The data was first cleaned, and then we conducted the statistical analysis. In particular, any data point with incomplete or flawed records of admission times, types of services needed, or departure times was eliminated. In addition, about 10 patients with irregularly long LOS ($>$ 6 months) were taken out of the consideration since we believe that these records are extreme and could deviate the actual LOS statistics. After the cleaning phase, 1193 out of 21268 records were eliminated, which leaves us with the total of 19075 data points.

2.6 Occupancy of the ICU

The occupancy of the ICU is determined by simulating the ICU with the exact admission and departure times of patients that are given from the data. The result gives only a lower estimate on the actual unit occupancy since patient records with flawed, incomplete, or extremely long staying time are not included in this calculation. Fig.2-2 shows the occupancy record of the ICU from 1998 to 2008. Examples of the occupancy in a single year are presented in Fig.2-3 for year 2001 and 2002, during which the unit capacity was 18 (including one crash bed).

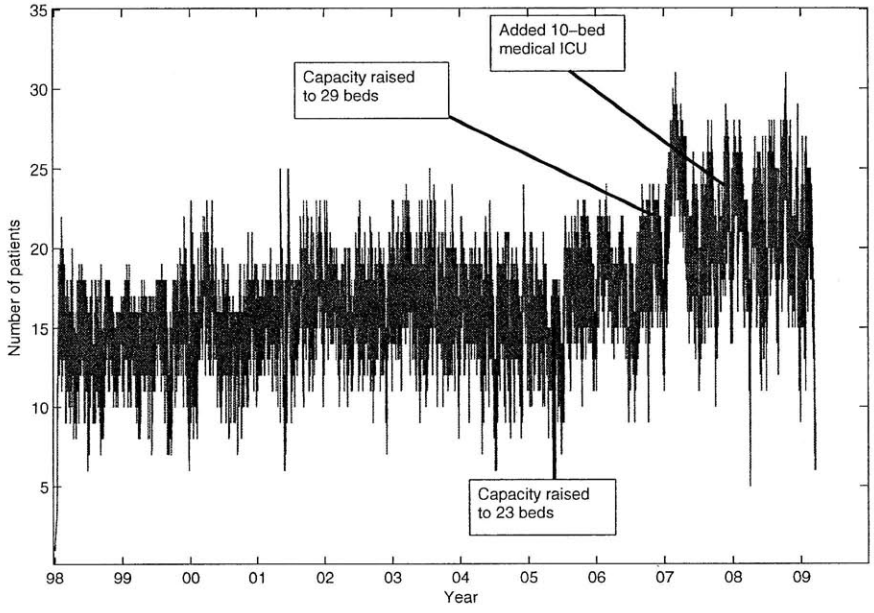
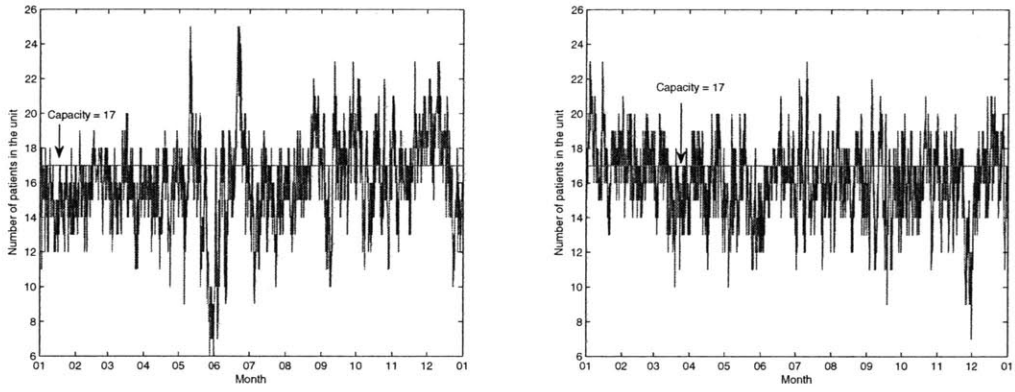


Figure 2-2: Occupancy of the ICU from 1998 to 2008

As can be seen in Fig.2-2, the evolution of the occupancy is consistent with the changes in the unit capacity. For example, the number of patients in the ICU increases in 2005 and 2007, which corresponds to the unit expansion in both of these years. We also observe the high utilization of the ICU during 1998 to 2005, which might have led to the decision to increase the capacity as well as to use admission control policies in the unit later on. Note that periods during which the unit occupancy exeeds the

capacity exist in both Fig.2-2 and 2-3 since the ICU data has the records of surgical patients who were diverted to stay in other units as well.



(a) 2001 (b) 2002

Figure 2-3: Occupancy of the ICU in 2001 and 2002 when the unit capacity was 18

2.7 Arrivals of ICU Patients

We analyze the statistics of ICU arrivals at different time scales, including years, seasons in the year, days of the week, and time of the day. The arrival data captures all the surgical patients that needed an ICU, but does not include medical patients who were not admitted by the ICU. That is, the data reflects the *admission* rate rather than the arrival rate of medical patients. In Section 3.2.2 of Chapter 3, we will discuss an approach to uncensor the true arrival rate of medical patients to the ICU and use that method to generate their arrivals in simulation.

2.7.1 Arrivals by Years

Table 2.2 and Fig.2-4 show the daily arrival rates of surgical patients and the daily admission rates of medical patients from 1998 to 2008. To reflect the actual number of arrivals/admissions, the arrival/admission rates are calculated from the data of all patients including those whose records are incomplete, flawed, or with irregularly

long LOS.

Year	Arrivals of surgical patients per day	Admissions of medical patients per day	Total arrivals per day
1998	2.90	1.70	4.60
1999	2.97	1.98	4.95
2000	3.32	1.77	5.09
2001	3.08	2.01	5.09
2002	3.09	2.09	5.18
2003	3.19	2.17	5.36
2004	2.99	1.96	4.95
2005	3.42	1.65	5.07
2006	3.36	1.82	5.18
2007	3.40	2.31	5.71
2008	3.89	1.88	5.77

Table 2.2: Daily arrival/admission rates of patients to the ICU from 1998 to 2008

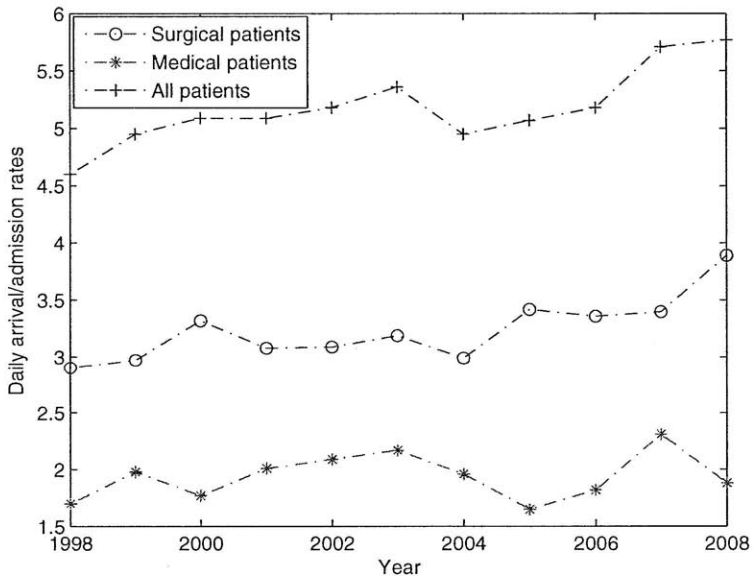


Figure 2-4: Daily arrival/admission rates of patients to the ICU from 1998 to 2008

Observe that the total arrival/admission rates tend to increase every year except for the drop in 2004. The sudden rise in 2007 is connected to the unit expansion from 23 to 29 beds.

2.7.2 Arrivals by Seasons

Year	Surgical patients' arrival rate per day		Medical patients' admission rate per day	
	Winter (December - March)	Non-winter (April - November)	Winter (December - March)	Non-winter (April - November)
1999	2.74	3.01	2.18	1.83
2000	3.15	3.34	2.38	1.52
2001	3.06	3.16	2.12	1.94
2002	3.04	3.09	2.20	1.92
2003	2.93	3.31	2.56	1.99
2004	2.51	3.18	2.40	1.79
2005	2.90	3.60	1.96	1.52
2006	3.32	3.47	1.69	1.79
2007	3.03	3.53	2.35	2.26
2008	3.67	4.15	2.57	1.53

Table 2.3: Seasonal arrival/admission rates per day of surgical and medical patients from 1999-2008

Surgical service	Winter daily arrival rate	Non-winter daily arrival rate
Neurosurgical	0.82	0.75
ORL[1]	0.51	0.55
Plastics	0.23	0.31
Urology	0.10	0.07
OMFS[2]	0.00	0.00
Orthopedic	0.49	0.64
Trauma	0.08	0.12
IntRadio[3]	0.08	0.13
General surgery	0.80	0.75
Other surgery	0.04	0.02
Sum	3.15	3.34

[1] Otorhinolaryngology
[2] Interventional radiology
[3] Oral and maxillofacial surgery

Table 2.4: Seasonal arrival rates per day of surgical patients by services in 2000

The daily arrival rates of surgical and the daily admission rates of medical patients during the winter (December to March) and the non-winter (April to November) seasons are presented in Table 2.3. The table clearly suggests the seasonal variation in

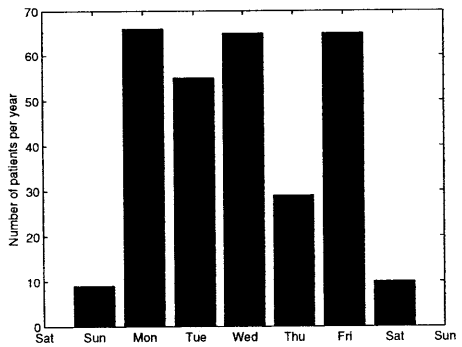
Surgical service	Winter daily arrival rate	Non-winter daily arrival rate
Neurosurgical	0.77	0.84
ORL	0.92	1.11
Plastics	0.13	0.26
Urology	0.20	0.03
OMFS	0.09	0.11
Orthopedic	0.46	0.55
Trauma	0.05	0.07
IntRadio	0.09	0.12
General surgery	0.91	0.99
Other surgery	0.05	0.07
Sum	3.67	4.15

Table 2.5: Seasonal arrival rates per day of surgical patients by services in 2008

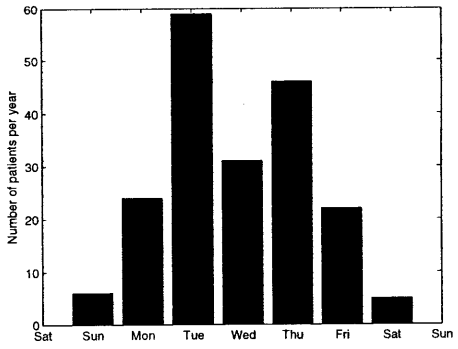
the arrival/admission rates. In particular, the demand for ICU from medical patients is higher in the winter. This could be a result of many kinds of seasonal illnesses, especially respiratory diseases, that are likely to spread during this season. Also notice from the table that the surgical arrival rates are lower in the winter months. This is probably because people are usually on vacations during the period from December to January. In addition, the seasonal arrival rates of surgical patients based on services in 2000 and 2008 are provided in Table 2.4 and Table 2.5, respectively. The arrival statistics of these two years are of particular interest since we will use them in conducting computational experiments later on in this thesis. As can be seen, surgical patients tend to arrive more during the non-winter season for most of the services. The seasonality of arrival rates will be taken into account when we develop a simulation model of the ICU in the next chapter.

2.7.3 Arrivals by Days of the Week

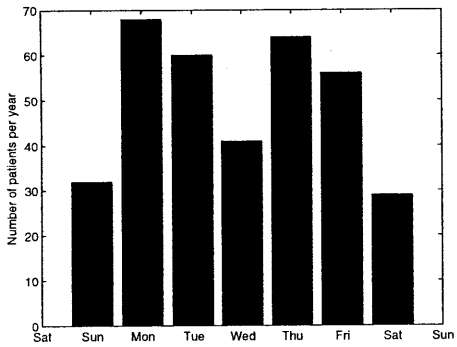
Fig.2-5 shows the total number of arrivals/admissions to the ICU by days of the week in 2008. It can be seen that the arrivals vary throughout the week. For surgical patients, the variation in arrivals is mostly a consequence of the block-based scheduling (see in Section 2.2), which varies on a daily basis. For example, the block time at CHB in 2008 allocates 2,1,2,1,2 ORs for neurosurgery from Monday through Friday,



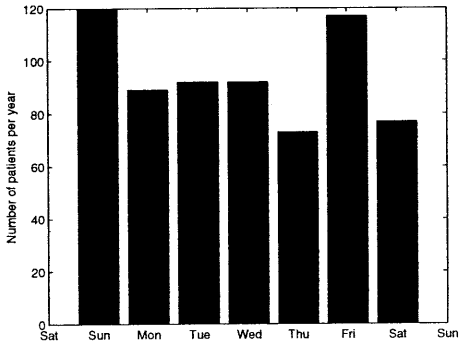
(a) Neurosurgical patients



(b) Orthopedic patients



(c) General surgery patients

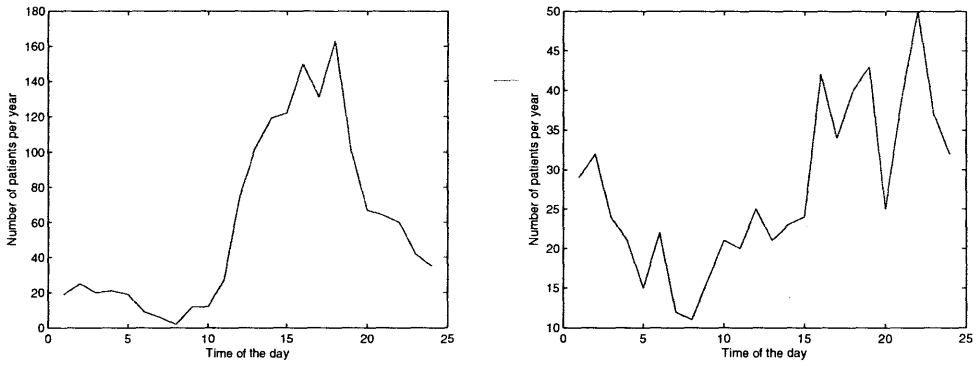


(d) Medical patients

Figure 2-5: Arrivals/admissions in each day of the week in 2008

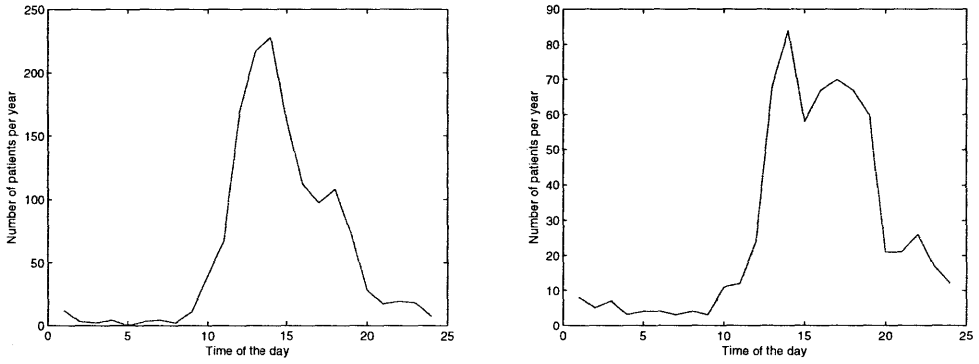
respectively and zero ORs over the weekend. The number of neurosurgical patient arrivals are therefore higher on Monday, Wednesday, and Friday than on the other days. Surgical arrivals on weekends are relatively few because they merely consist of emergency patients. Note that we show these arrival statistics just for year 2008 since we only have the block information in 2008.

In contrast, medical patients are not scheduled and tend to arrive randomly over the week. Moreover, unlike surgical patients, many medical patients can arrive on weekends as evidenced by the high number of medical admissions on Sunday in 2008.



(a) Surgical patients (b) Medical patients

Figure 2-6: Admission times of patients in 2008. The horizontal axis represents the time of the day from 1 AM to 12 AM.



(a) Surgical patients (b) Medical patients

Figure 2-7: Departure times of patients in 2008. The horizontal axis represents the time of the day from 1 AM to 12 AM.

2.7.4 Arrivals by Time of the Day

The admission times of surgical and medical patients in 2008 are shown in Fig.2-6. As can be seen, a large portion of surgical patients enter the ICU in the afternoon and the evening. This finding is unsurprising because most scheduled surgical patients arrive to the OR in the morning, undergo operations, and are transferred to the ICU in the afternoon. On the other hand, medical patients seem to come to the ICU in a more random pattern compared to surgical patients, even though the figure suggests

that their arrivals are more rare in the early morning and higher in the afternoon and the evening.

Fig.2-7 displays the departure times of surgical and medical patients in 2008. Notice that most patients leave the ICU in the afternoon a few hours before the admissions of surgical patients. This is because the ICU usually discharges current ready-to-leave patients in order to make beds available for newcoming surgical patients a few hours prior to their arrivals.

2.8 Length of Stay

2.8.1 Length of Stay by Years

Year	Surgical patients			Medical patients		
	Mean (hours)	SD (hours)	CV	Mean (hours)	SD (hours)	CV
1998	64.89	122.15	1.88	106.66	184.97	1.73
1999	72.39	137.82	1.90	92.46	136.11	1.47
2000	66.20	142.38	2.15	121.85	257.46	2.11
2001	73.44	120.43	1.64	108.07	201.81	1.87
2002	73.46	139.25	1.90	97.10	189.91	1.96
2003	73.35	133.37	1.82	94.02	173.42	1.84
2004	65.92	120.80	1.83	109.73	184.74	1.68
2005	71.88	150.11	2.09	122.33	230.24	1.88
2006	85.20	187.54	2.20	119.77	221.39	1.85
2007	71.87	135.71	1.89	124.91	233.24	1.87
2008	76.43	181.37	2.37	120.86	219.91	1.82

Table 2.6: Mean, SD, and CV of ICU patients' LOS from 1998 to 2008

Table 2.6 shows the mean, SD, and coefficient of variation (CV: the SD divided by the mean) of the LOS of surgical and medical patients from 1998 to 2008. Notice that the mean and the CV fluctuate yearly throughout the decade. This is mainly due to the variation in the severity and chronicity of patients' diseases, as well as the fundamental change in patients' services and population from year to year.

In addition, the LOS of ICU patients are highly variable as can be seen from the high SD of both types of patients. The table also suggests that medical patients'

LOS are significantly longer on average, but with roughly the same CV as surgical patients' LOS. However, we will see that some types of surgical patients are similar to medical patients in terms of their LOS statistics when we investigate the LOS of each type of service in Section 2.8.3.

See Table A.1 and Table A.2 in Appendix A for further information about the LOS statistics based on patients' services from 1998 - 2008.

2.8.2 Length of Stay by Seasons

Year	Mean LOS of surgical patients (hours)		Mean LOS of medical patients (hours)	
	Winter	Non-winter	Winter	Non-winter
1999	79.63	72.67	91.92	99.97
2000	60.95	66.07	104.94	125.03
2001	73.44	78.21	108.95	101.76
2002	69.67	75.10	110.06	99.33
2003	66.40	76.63	103.37	87.76
2004	63.72	71.66	107.31	105.16
2005	81.67	69.94	112.39	131.07
2006	108.26	84.07	138.68	116.30
2007	76.81	70.65	141.12	109.94
2008	69.24	79.00	107.57	129.91

Table 2.7: Average LOS of the ICU patients in the winter and non-winter seasons from 1998 to 2008

Table 2.7 shows the mean of the LOS by seasons from 1998 to 2008. According to medical experts, the LOS of medical patients can be dependent on seasons due to seasonal diseases. For example, medical patients with respiratory illnesses such as influenza and respiratory syncytial virus (RSV) are more common in the winter and usually need more time in the recovery process. Nonetheless, the severity of respiratory symptoms depends on years, which means that there are years (e.g., 1999, 2000, 2005, and 2008) in which these diseases might not be as serious and the LOS in the winter of those years could be lower on average. According to medical professionals, the LOS of surgical patients can also be seasonally dependent. This could be due to the difference between portfolios of surgical patients in the winter and non-winter

seasons.

2.8.3 Length of Stay by Services

The LOS statistics of the admitted patients is crucial to the understanding of availability in the ICU. We believe that long-stay patients are one of the major causes for service rejections in the ICU. Table 2.8 shows the mean and the SD of the LOS by types of services from 1998 to 2008.

Type	Subtype	Percentage of patients	Mean of LOS (hours)	SD of LOS (hours)
Surgical	Neurosurgical	13.67%	51.35	88.50
	General surgery	13.40%	121.08	239.95
	Orthopedic	7.74%	78.90	130.16
	ORL	10.98%	52.41	90.18
	Plastics surgery	4.19%	50.49	62.17
	Trauma	1.93%	61.55	107.33
	Urology	0.95%	55.97	58.64
	IntRadio	1.62%	68.09	83.08
	OMFS	1.18%	51.00	63.85
	Other surgery	0.71%	79.84	206.39
Medical	-	40.45%	100.10	203.87

Table 2.8: Mean and SD of the LOS from 1998 to 2008 based on types of services

According to the table, we can classify surgical patients into three groups based on the mean and SD of their LOS as follows.

- Group 1, which consists of surgical services that have short LOS on average ($\sim$ 50 hours). Constituting 5911 cases (30.97% of all patients), this group includes neurosurgical, ORL, plastics surgery, urology, and OMFS patients.

- Group 2, which consists of surgical services that have intermediate LOS on average (60 - 80 hours). Constituting 2155 cases (11.29% of all patients), this group includes orthopedic, trauma, and IntRadio patients.

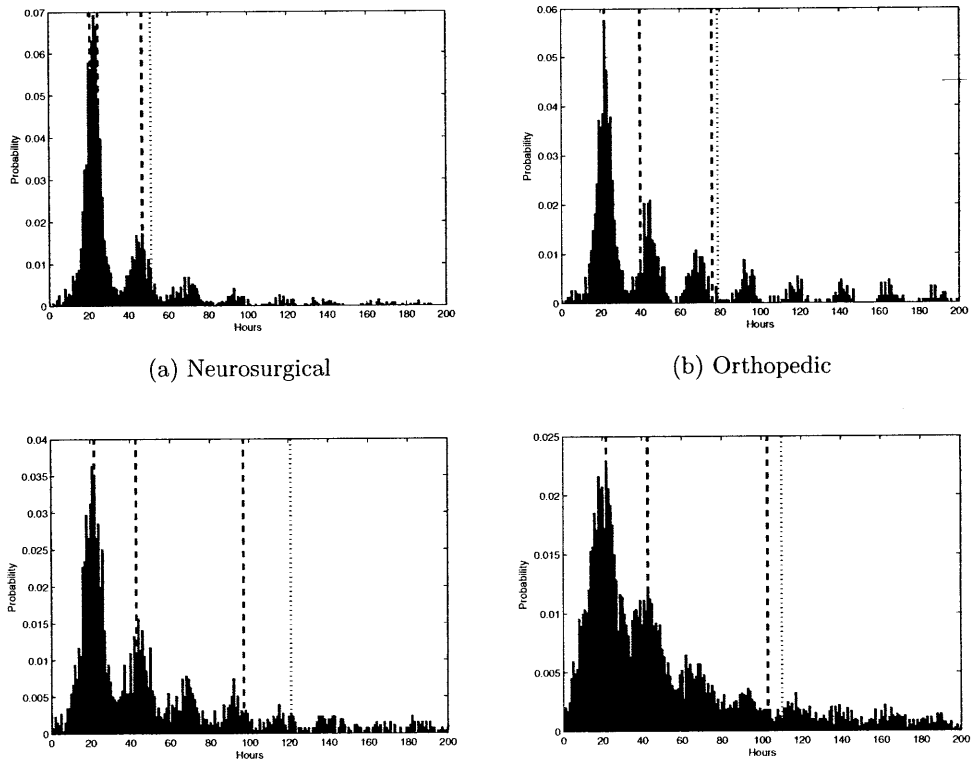
- Group 3, which consists of surgical services with LOS that are long on average ($>$ 100 hours) or highly variable. Constituting 2691 cases (14.11% of all pa-

tients), this group includes general surgery and other surgery patients. We keep other surgery patients in this group because their LOS, although much shorter compared to general surgery patients', are highly variable. It should be noted that general surgery patients tend to have longer stays due to the complicated nature of the underlying disease processes rather than the nature of surgery. For example, congenital diaphragmatic hernia (CDH) cases involve relatively simple surgery but a very challenging pathophysiology.

Observe that the LOS of Group 3 is significantly longer on average and more variable compared to the other two groups of surgical patients. In fact, Group 3 of surgical patients and medical patients share the similar statistics of the LOS. Both are likely to stay a very long period in the ICU, and could block the ICU for a number of days if many of them are allowed to enter the unit at the same time. This is where the SSCP is expected to play a role in controlling the continual influx of the long-stay types of patients into the ICU, which could help improving the overall patient flow in the unit. We will discuss in detail the scheduling mechanism of the SSCP as well as its performance in Chapter 4

Fig.2-8 shows the distributions of the LOS based on the data from 1998 to 2008 of several types of patients. As can be seen, neurosurgical patients are likely to leave after a few days of stay, while orthopedic, general surgery, and medical patients tend to stay longer. The LOS distributions of patients with other surgical services are provided in Fig.A-1 of Appendix A.

It should be noted from the Fig.2-8 that surgical patients tend to stay in the ICU for the full-day lengths, e.g., at 24, 48, 72 hours. This result is consistent with the admission and departure times of surgical patients, both of which frequently occur in the afternoon (see Fig.2-6 and Fig.2-7). On the other hand, the LOS of medical patients tend to spread more. This finding is as well consistent with the admission and departure times of medical patients, as they tend to come to the ICU randomly during the day, but leave mostly in the afternoon (again, see Fig.2-6 and Fig.2-7).



(a) Neurosurgical

(b) Orthopedic

(c) General surgery

(d) Medical

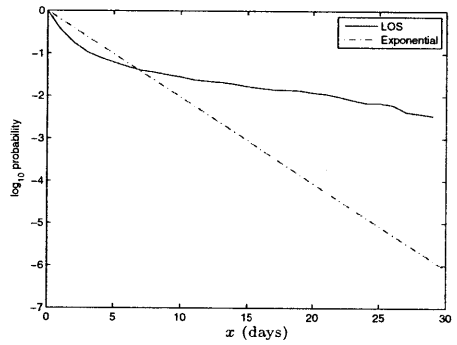
Figure 2-8: Distributions of the LOS from 1998-2008. The first three dashed lines indicate the first, second, and third quantiles respectively, while the dotted line locates the mean of the LOS. The horizontal axis is truncated to 200 hours.

2.8.4 Tail Distributions of the Length of Stay

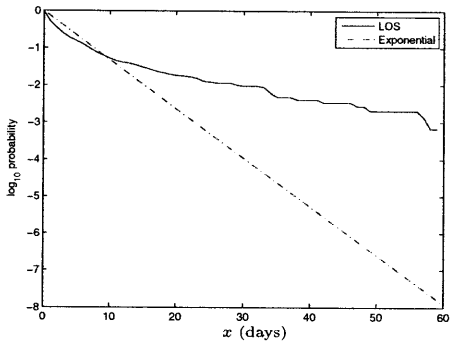
We now investigate the tail distributions of the LOS, namely $\Pr(\text{LOS} > x)$ by comparing them with certain parametric distributions.

Exponential Distributions

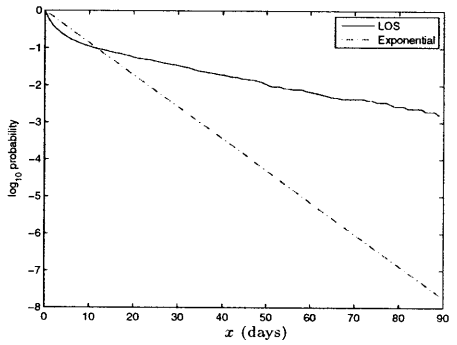
We are interested in fitting the tail distributions of the LOS with those of exponential distributions since exponentially-distributed service times facilitate the analysis of many queueing systems and could provide explicit results in terms of simple closed-form formulas (Kleinrock [20]). Let X be an exponential random variable with mean



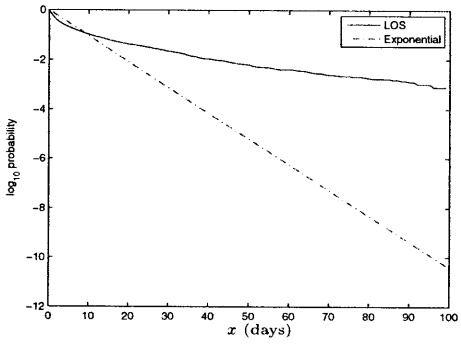
(a) Neurosurgical ($\lambda = 2.14$)



(b) Orthopedic ($\lambda = 3.29$)



(c) General surgery ($\lambda = 5.05$)



(d) Medical ($\lambda = 4.17$)

Figure 2-9: Tail distributions of the LOS from 1998-2008 and of exponential distributions with mean corresponding to the average LOS of each service

$1/\lambda$. The cumulative distribution function (CDF) of X is given by

$$F(x;\lambda) = 1 - e^{-\lambda x},\ x \geq 0.$$

The tail distribution of X is therefore equal to $\Pr(X > x;\lambda) = 1 - F(x;\lambda) = e^{-\lambda x}$ for $x \geq 0$. Equivalently, $\log \Pr(X > x;\lambda) = -\lambda x$, $x \geq 0$.

Fig.2-9 illustrates the LOS tail distributions of several services on the same plot as the tails of exponential distributions with mean corresponding to the average LOS of those services. The graph is given as a log plot so that the tails of exponential distributions are presented as straight lines, which are easy to compare them with those of the LOS distributions. As can be seen, the tail distributions of the LOS do

not match the tails of exponential distributions. In fact, the LOS tail distributions start off falling faster up to some point after which their rates of decay become slower compared to those of the exponential tail distributions.

Weibull and Log-normal Distributions

We further explore the tail distributions of the LOS with respect to the tails of the Weibull and log-normal distributions, both of which are widely used in the parametric modeling of LOS distributions as well as lifetime data (Lawless [25] and Marazzi et al. [28]).

Let X be a Weibull random variable parametrized by $\alpha > 0$ and $\beta > 0$. The CDF of X is given by (Lawless [25])

$$F(x; \alpha, \beta) = 1 - e^{-(x/\alpha)^\beta}, \ x \geq 0,$$

which simply implies that the tail distribution of X is equal to $\Pr(X > x; \alpha, \beta) = e^{-(x/\alpha)^\beta}$, $x \geq 0$. Consequently, $\log \Pr(X > x; \alpha, \beta) = -(x/\alpha)^\beta$ for $x \geq 0$. Notice that when $\beta = 1$, X becomes an exponential random variable with mean α. Fig.2-10 presents the log plots of the LOS and Weibull tail distributions of several services. The horizontal axis is scaled to $(x/\alpha)^\beta$ so that the tails of Weibull distributions are expressed as straight lines, which are convenient for us to compare them with those of the LOS distributions. The parameters α and β are varied in order to fit the tail of the Weibull distribution to the tail LOS distribution of each service.

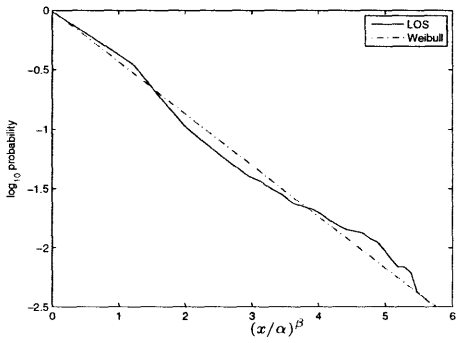
A random variable X is said to be log-normally distributed if $Y = \log X$ is normally distributed with mean μ and variance σ^2 (Lawless [25]). Its probability density function (PDF) is given by

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, \ x > 0.$$

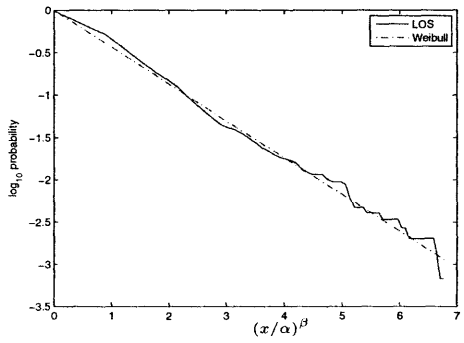
Since the logarithm of this PDF is dominated by $\log^2 x$ function, we present its tail distribution in a log plot with the horizontal axis scaled to $\log^2 x$. Fig.2-11 shows

the tail distributions of the LOS and log-normal random variables. In each plot, the parameters μ and σ are adjusted in order to match the log-normal tail distribution with the corresponding tail distribution of the LOS.

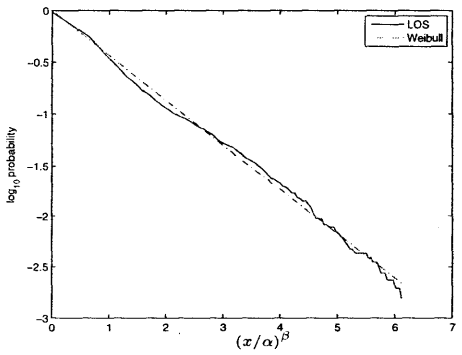
Both Fig.2-10 and 2-11 show that Weibull and log-normal distributions can be appropriately used as an parametric estimation for the tail distributions of the LOS, even though the approximation from Weibull distributions seem to give a better fit. As a result, we conclude that both Weibull and log-normal random variables are suitable candidates to model the LOS of the ICU patients at CHB.



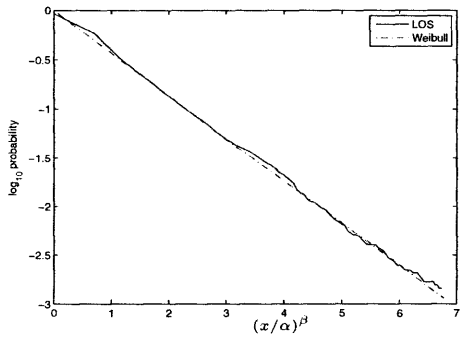
(a) Neurosurgical ($\alpha = 0.67$, $\beta = 0.46$)



(b) Orthopedic ($\alpha = 1.29$, $\beta = 0.5$)

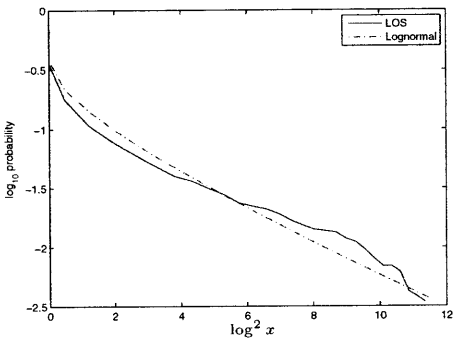


(c) General surgery ($\alpha = 2.38$, $\beta = 0.5$)

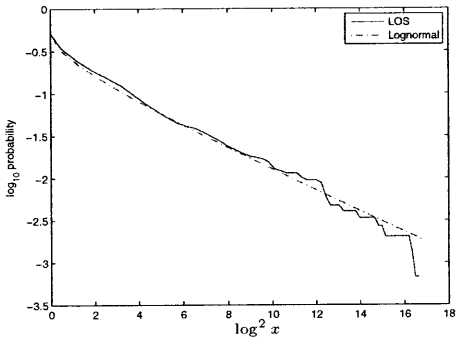


(d) Medical ($\alpha = 1.96$, $\beta = 0.5$)

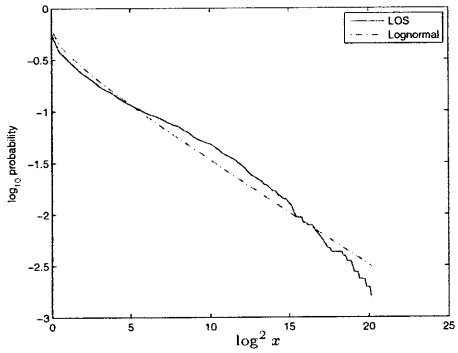
Figure 2-10: Tail distributions of the LOS from 1998-2008 and of the Weibull random variables



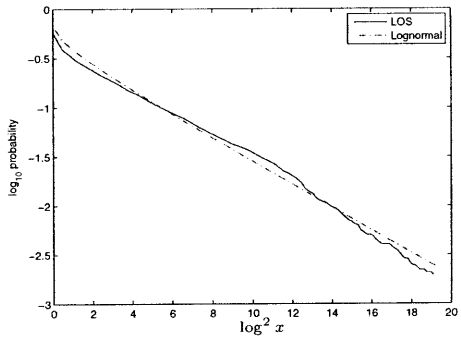
(a) Neurosurgical ($\mu = -0.42$, $\sigma = 1.42$)



(b) Orthopedic ($\mu = 0.04$, $\sigma = 1.40$)



(c) General surgery ($\mu = 0.5$, $\sigma = 1.46$)



(d) Medical ($\mu = 0.63$, $\sigma = 1.33$)

Figure 2-11: Tail distributions of the LOS from 1998-2008 and of the log-normal random variables

Chapter 3

Simulation Model

In this chapter, we describe a simulation model which captures the main dynamic aspects of the behavior of the ICU system at Children's Hospital Boston (CHB) as it evolves over time. The goal of the simulation-based model is to evaluate the performance of various policies in the ICU. All the computational experiments and simulations were implemented using MATLAB. The simulation model is calibrated based on the specific ICU environment and data discussed in Chapter 2. We validate the model by simulating the system with the ICU data in year 2000, where all arrivals including rejected patients were documented, and then show that the simulation provides results consistent with the actual data. In particular, our model is verified to provide accurate estimates for the rejection rates and the system utilization of the real ICU in 2000

Next, we describe the modeling framework and discuss the underlying assumptions.

3.1 General Framework

The simulation model of the ICU at CHB is constructed based on the discrete-event simulation framework. Time evolution in the model is continuous and corresponds to the actual time clock and calendar. The input to the model is two streams of arrival epochs, namely scheduled (elective) and emergency patients. These two types

of patients can consist of several subtypes, each of which is generated separately as a stochastic process with a time-varying rate. Specifically, the arrivals to the ICU of each stream is represented by a sequence of arrival times, whereas the time between consecutive arrivals (the interarrival time) is independently drawn from a probability distribution. According to medical experts, since the majority of emergency patients require care related to non-surgical issues, we assume that all emergency arrivals are medical patients. Thus, it is immediate that all scheduled patients are surgical patients.

Surgical patients arrive to the queue and are scheduled for surgery on certain dates as per the implemented scheduling policy, while medical patients arrive to the ICU directly. Upon arrivals to the ICU, all patients have to go through an admission process, which decides whether or not their requests for admissions are accepted. In this thesis, we restrict our attention to stationary policies that do not use the current state of the ICU to make decisions. Nevertheless, we want to emphasize that our simulation model can incorporate any predefined set of admission and scheduling policies.

The lengths of stay (LOS) of ICU patients are drawn from the respective empirical distribution of LOS data according to their services and seasonality. This is one particular choice we select for modeling the LOS, although in fact we can generate the LOS from any kind of distributions. Patients stay in the unit and leave when their LOS expire. Once patients depart, those previously-occupied beds are immediately free and ready to take on new patients at once.

3.2 The ICU at CHB

In this section, we describe in detail the simulation model that incorporates the specifics of the ICU at CHB.

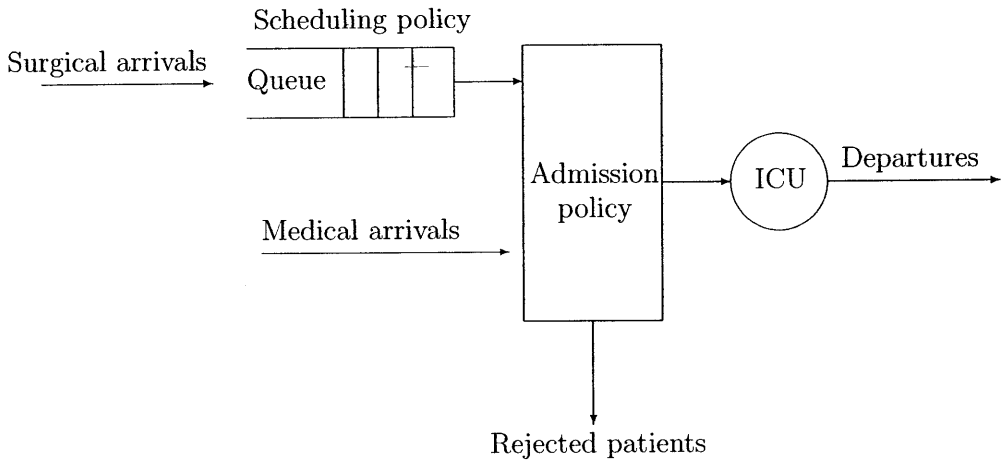


Figure 3-1: Simulation diagram

3.2.1 Time

Time horizon in the simulation comprises two seasons: the winter, which ranges from December to March, and the non-winter, which covers the rest period of the year.

3.2.2 Arrivals

As described in the previous section, there are two main streams of arrivals to the ICU in our model: surgical and medical. Surgical arrivals include ten types of services as listed in Table 3.1. Each of them is assumed to arrive according to a non-homogeneous Poisson process with a seasonally-varying rate. The seasonal arrival rate of each surgical service is estimated by averaging the number of surgical arrivals per day of the corresponding type and season based on the data in the years that we want to simulate the system. We set surgical arrival rates to be constant through each day of the week (excluding the weekends, over which the surgical departments are close) because the data does not provide the dates on which surgical patients arrived to the OR booking office. Indeed, our reason for setting surgical arrival rates to be constant is the fact that the block times are fixed throughout. Varying arrival rates

by days of the week would not make significant difference regarding the number of scheduled surgical patients that enter the ICU on each day since their schedules are controlled by surgeons' fixed block times. We will see that, even with this simplifying assumption, we are able to obtain accurate estimates.

The arrivals of medical patients to the ICU are also characterized by a non-homogeneous Poisson process with a rate depending on the day of the week and the season. The *admission* rate is computed as the mean number of medical admissions on the corresponding arrival day and season from the actual data in the years that we want to simulate the system. We will discuss below how to estimate the actual *arrival* rate of medical patients from the admission rate. Unlike surgical patients, all medical patients belong to the emergency class and therefore come to the ICU directly without prior scheduling.

Calculating Arrival Rates for Medical Patients

As the data of medical patients represents only those who were admitted (except for year 2000 which has an accurate documentation of the number of rejected medical patients), we describe a method to uncensor their true arrival rate. On day i of the week in a given season, let A_i' be a corresponding medical *admission* rate from the data and A_i be a corresponding *arrival* rate. We assume that all medical patients are rejected when they come to the fully-occupied ICU. Moreover, let T be the time interval over which we want to compute the arrival rate, and let T_B be the amount of time that the ICU is fully occupied in the interval T. Since the average arrival rate is assumed to be constant on any day i of the week during the interval T, it follows that the mean number of admissions in T, $A_i'T$, is equal to the mean number of arrivals in T but outside T_B, $A_i(T - T_B)$. Therefore,

$$\frac{T - T_B}{T} = \frac{A_i'}{A_i}, \tag{3.1}$$

and the *uncensored* medical arrival rate A_i is obtained.

We now use the data on medical rejections from year 2000 to validate this uncen-

soring method. The ratio of medical rejections in 2000 is 0.13, while the rejection rate (namely the arrival rate subtracted by the admission rate) of medical patients calculated from our method is 0.15. This overestimate is expected since our uncensoring method assumes that patients are rejected if they arrive during the fraction of time when the unit is full. However, in such an interval, it could be the case that referral hospitals did not call the ICU since they knew in advance that the unit was already full. While the number of medical rejections counted by the ICU never included these "hidden" rejections, our uncensoring method takes them into account. Nonetheless, this overestimating gap is not so significant, so we will use this method to uncensor medical arrival rates.

3.2.3 Scheduling Policy

The block-based scheduling is used as the baseline policy to schedule elective patients in the simulation model. Our scheduling policy assumes that only one elective case that requires an ICU can be scheduled to one open OR per day. This is because we do not know the arrival rate of "all" surgical patients (which include both ICU and non-ICU patients) to the OR department and how many cases one OR can operate per day on average. In addition, since we have no access to the block of each individual surgeon, our simulation assumes that elective patients do not select surgeons based on their preference, but are always scheduled to the very first open OR that is available.

Following these assumptions, all elective patients who arrive to the queue are scheduled to the first available slot by the First-In-First-Out (FIFO) discipline according to their surgical services. If there are caps enforced, a patient that requires a post-operative ICU bed may not be scheduled into the first available block because of the caps. In this case, our assumption is that a non-ICU patient will be scheduled to fill that block instead. In this chapter, we focus merely on the block-based scheduling policy. The next chapter will consider the uniform cap and the service-specific cap policies, which we introduced in Section 2.2 of Chapter 2.

It is important to note that, in reality, elective patients always request specific surgeons upon their booking processes, and their surgery dates will be limited to the

available block of the surgeons of their choices. Since our simulation assumes that elective patients are scheduled to the very first available block regardless of surgeons (given caps permitted), the corresponding mean waiting time would likely be shorter than it should be in the actual surgeon-based scheduling system. In addition, the assumption that any ICU elective case will always be scheduled to the first available open OR might not apply to the real situation, and could as well affect an estimate for mean waiting time from simulation. Since most surgeries do not require an ICU, it is possible in the real system that the most recent available OR has been completely filled by all non-ICU cases, which would delay the scheduling of an ICU patient. As such, our simulation is also likely to give a relatively shorter mean waiting time of scheduled surgical patients compared to the real one in this case. Nonetheless, we will show in Section 3.3.2 that the simulation built on this scheduling process is able to provide accurate estimates for various performance metrics of the ICU.

3.2.4 Admission Policy

As described in Section 2.3 of Chapter 2, the admission policy in the actual ICU gives priority to medical patients, since they are usually sicker and can be cared only in the ICU. When a medical patient arrives to a full unit, often times, the ICU will move its current surgical patients to other units in order to generate a free bed to admit this medical patient. To capture this policy, we model a *secondary* ICU (SICU), with limited capacity, to be a medical unit that treats surgical patients who are diverted from the main ICU. It is important to note that the SICU in the simulation model is not restricted to one single back-up location in reality, but it can represent multiple overflow units for caring surgical patients who are diverted from a full ICU. Examples of such units in CHB include the post anesthesia care unit (PACU) and the cardiac ICU. By incorporating this SICU into the model, the admission rules for surgical and medical patients in our simulation are as follows.

- Any patient who arrives to an ICU that is not fully-occupied is always admitted.
- When a surgical patient arrives to a full ICU,

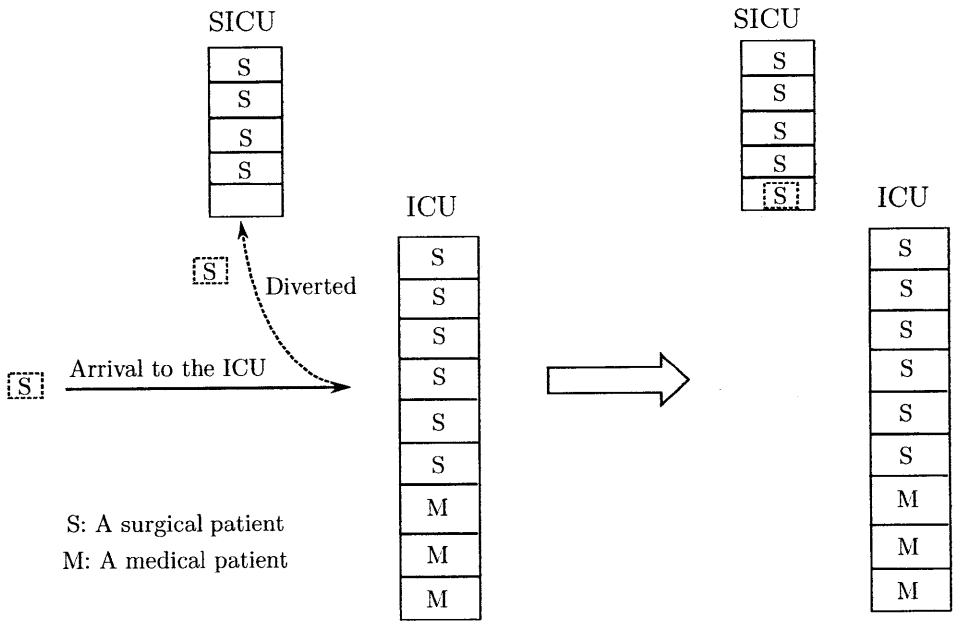


Figure 3-2: The illustration of the admission process for surgical patients when the ICU is full. In this case, an arriving surgical patient is diverted to the SICU. If the SICU is full, this surgical patient will be rejected from the system

  - if there is space within the SICU, divert this patient to this unit.
  - Otherwise, if the SICU is full, reject this surgical patient from the system.

- When a medical patient arrives to a full ICU,
  - if there exist current surgical patients in the ICU, divert one of them to the SICU if space permitted; otherwise, simply reject this medical patient from the system.
  - If there is no surgical patient currently occupying the ICU, reject this medical patient from the system.

Regarding the diversion of surgical patients to provide space for a medical patient, we note that it is actually done on a case-by-case basis according to the current health condition of each surgical patient, and it is hard to model this process exactly. Since there are no precise rules on the diversion procedure, we decide to *uniformly* select

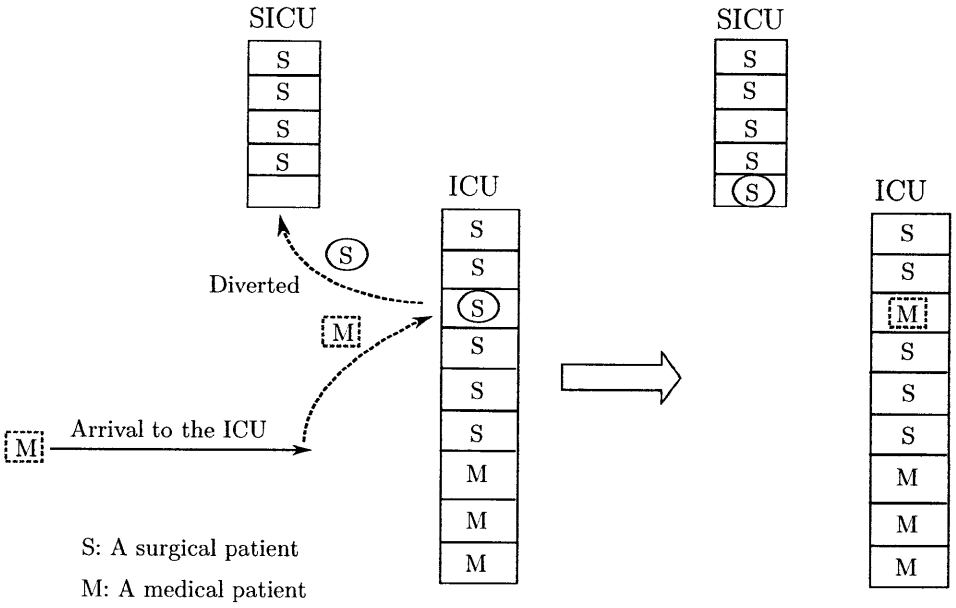


Figure 3-3: The illustration of the admission process for medical patients when the ICU is full. In this case, the ICU uniformly selects one of its surgical patients and move him to be cared in the SICU so as to provide the space for an arriving medical patient. This medical patient will be rejected when either the SICU is full or there is no single surgical patient to be diverted from the main ICU.

one of the current surgical patients when one needs to be diverted. We shall see later that, even with this simplifying assumption, we are able to obtain relatively accurate estimates for performance measures of surgical and medical patients.

It is important to note that, in reality, there cannot be rejections for ICU surgical patients once their surgeries have started (see Section 2.3 in Chapter 2). In case that the unit is full, the ICU is obligated to find space in other medical units to treat surgical patients post-operatively. For this reason, the rejections of surgical patients in our simulation are interpreted as the diversions from the ICU to another location.

With this admission policy, the more space provided to the SICU, the more priority is given to medical patients. Therefore, we expect a smaller number of medical rejections and, at the same time, a higher number of surgical diversions/rejections when the SICU capacity is raised.

Remarks on the 10-bed medical ICU

In Chapter 6, we will simulate the ICU scenario in year 2008, which includes the new 10-bed medical ICU. As discussed in the previous chapter, there are no clear rules regarding which unit medical patients should be sent to. Moreover, at this point we still have no access to the database of the new unit. These difficulties prevent us to appropriately model the 10-med medical ICU, and we will not incorporate it into the simulation model of the ICU system.

3.2.5 Event Timelines of Surgical and Medical Patients

Surgical Patients

Consider a surgical patient who arrives to the queue at time T_1. There, he is given a surgery date, which is at time T_2, via a scheduling procedure. His waiting time in the queue is therefore equal to $W_{queue} = T_2 - T_1$. At time T_2, he leaves the queue and waits until T_3 to enter the ICU. The waiting time period outside the queue, denoted by $W_{OR} = T_3 - T_2$, can be viewed as the operation time. W_{OR} is generated from the empirical distribution of the admission time of the day data that corresponds to the type of service and the season in which this request for admission is made. The method of generating a random value from an empirical distribution is described in Appendix B. Note that only W_{queue}, not W_{OR}, will represent the waiting time of each scheduled surgical patient in the queue.

At time T_3, the patient arrives to the ICU and encounters the admission process. If he sees a full ICU, he is diverted to the SICU if space permitted. Otherwise, he is rejected from the SICU, in which case he leaves the system. On the other hand, if the admission is made at the ICU, this patient enters the unit and stays until his LOS expires. Similar to the sampling of W_{OR}, the LOS is drawn from the respective empirical distribution of the LOS data according to the surgical service and the season in which he arrives to the system. The patient leaves the ICU (or the SICU in case he was diverted at T_3) at time $T_4 = T_3 + \mathrm{LOS}$, after which a new bed is immediately available to serve the next patient in the ICU (or the SICU).

According to our admission rules, it is also possible that this surgical patient is admitted, but then the arrival of a medical patient at T_3' forces him to be diverted to the SICU. This occurs given that the main ICU is full and there is space available in the SICU at T_3'. In this case, the surgical patient is moved to the SICU, where he is cared and continues to stay there until his original LOS expires at T_4.

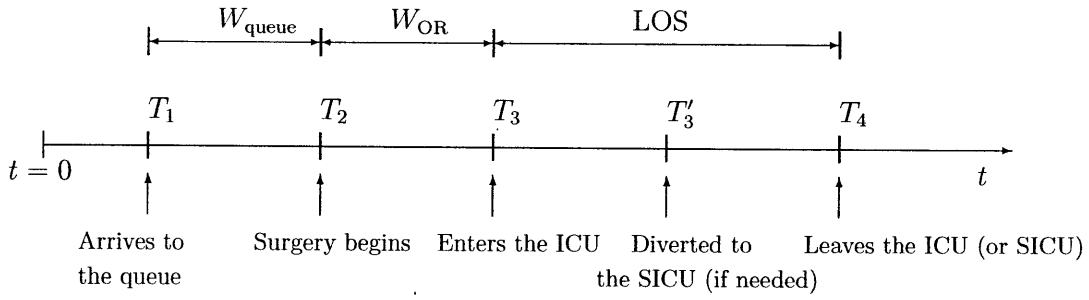


Figure 3-4: The event timeline of a surgical patient who is admitted to the ICU. In case that he needs to be diverted at time T_3', he stays in the SICU until his original LOS expires at time T_4.

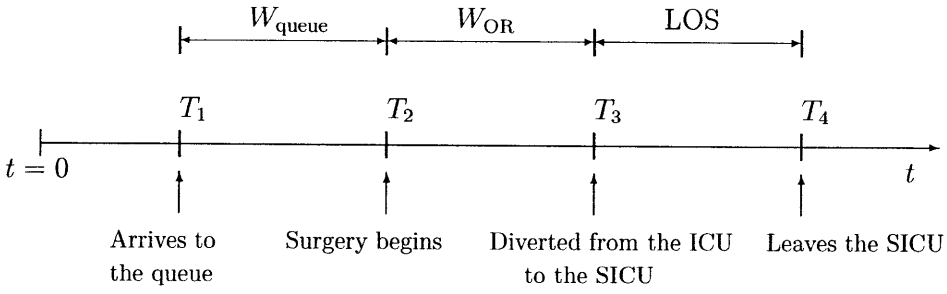


Figure 3-5: The event timeline of a surgical patient who is diverted to the SICU. This patient will be rejected at time T_3 if the SICU is full.

Medical Patients

Consider a medical patient who arrives to the ICU and goes through the admission process at time T_1. If he sees a full ICU and the situation is that either no surgical patients can be moved to the SICU because it is also full or no single surgical

patient currently exists in this full ICU, this medical patient is rejected and leaves the system immediately. Otherwise, he is admitted and stays for the duration of his LOS, which is drawn from the empirical distribution of the medical patients' LOS data corresponding to the season in which he arrives. Then, he departs as soon as the LOS expires, leaving his bed empty at time T_2 to serve a newcoming patient to the ICU.

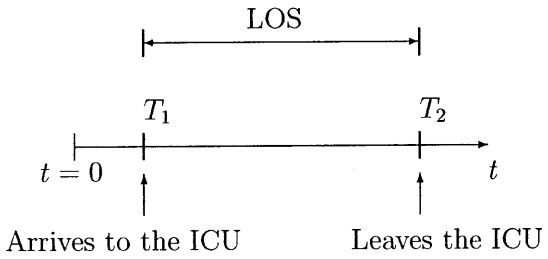


Figure 3-6: The event timeline of an admitted medical patient. A rejected medical patient leaves the system at time T_1.

3.3 Validation of the Simulation Model

To verify the simulation model, we simulate the system based on the data and ICU environment in year 2000, and then compare the rejections rates and utilization obtained from the simulation with those from the data in 2000. We choose year 2000 because the data from this year provides the complete number of arrivals to the ICU including the medical arrivals who were rejected.

3.3.1 Simulation Scenario

Surgical and medical patients arrive to the system according to the independent Poisson processes with seasonally-varying rates provided in Tables 3.1 and 3.2. The winter arrival rates are calculated according to the arrivals in January to March and December of 2000, while the non-winter arrival rates are computed from the arrivals in the rest period of the year. Note that the rates of surgical arrivals provided in

Table 3.1 are computed to be proportional to the number of weekdays in 2000, and they are higher than the rates in Table 2.4 of Chapter 2, which are weighted by the total number of days in 2000. Surgical patients are scheduled by the block-based policy with no caps. The number of beds is set to be 16 in the simulation since the average number of beds in 2000 was 17 including a crash bed. LOS and W_{OR} are generated from the associated empirical distributions of the data from 1998 to 2003, but not from year 2000 in order to prevent the data-snooping bias. We choose not to use the training data from 2004 to 2008 because the LOS after 2004 tend to be longer on average than those before 2004, which are unrealistic to use for validating the model in year 2000. The only parameter left to be defined is the capacity of the SICU. This will be discussed in detail in the next section.

Surgical service	Winter daily arrival rate	Non-winter daily arrival rate
Neurosurgical	1.08	1.07
ORL	0.77	0.74
Plastics	0.32	0.45
Urology	0.11	0.10
OMFS	0.00	0.00
Orthopedic	0.67	0.94
Trauma	0.16	0.15
IntRadio	0.15	0.18
General surgery	1.11	1.04
Other surgery	0.05	0.03
Sum	4.42	4.70

Table 3.1: Arrival rates per day of surgical patients in 2000

Season	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Winter	2.54	2.85	2.69	2.69	2.54	3.08	2.85
Non-winter	2.07	1.73	1.84	1.43	1.79	1.82	1.90

Table 3.2: Arrival rates per day of medical patients in 2000

3.3.2 Results and Discussion

100 sample paths of patients are generated based on the scenario described in the previous section. Each sample path contains the records of arrival times, W_{OR}, and

LOS. Moreover, it is generated for the period of two years, whereby the first year allows the system to reach steady state and the performance measures associated with each sample path are computed from its trajectory in the second year only. We are interested in comparing the rejection rates and utilization level obtained from the actual data in year 2000 and from simulation.

The rejections of surgical and medical patients are defined as follows.

- Surgical rejections
  - Surgical rejections from the data are computed from the actual number of surgical patients who were diverted to other units.
  - Surgical rejections from simulation are computed from the number of surgical patients who are either rejected or diverted to the SICU in the simulated system.
- Medical rejections
  - Medical rejections from the data are computed from the actual number of medical patients who were rejected from the ICU.
  - Medical rejections from simulation are computed from the number of medical patients who are rejected from the simulated system.

For each service, the seasonal rejection rates are calculated as a ratio between the number of seasonal rejections and the total number of seasonal arrivals, while the total rejection rates are computed from weighting the total number of rejections by the total number of arrivals. Finally, the mean rejection rates of surgical and medical patients from simulation are computed by averaging the rejection rates obtained from each sample path.

The utilization of the ICU in 2000 is computed from the records of patients in 2000 who were admitted to the main ICU only. This is to prevent the overestimate of the true ICU utilization by including those who were in fact admitted to the the cardiac ICU or the PACU. Out of the total 1377 patients that entered the main ICU in

Parameter	Result from year 2000	Simulation Result
Winter surgical rejection rate	41.98%	34.06%
Winter medical rejection rate	11.58%	27.63%
Total winter rejection rate	**28.55%**	**31.15%**
Non-winter surgical rejection rate	33.45%	27.63%
Non-winter medical rejection rate	13.86%	21.59%
Total non-winter rejection rate	**26.69%**	**25.56%**
Surgical rejection rate	36.21%	29.98%
Medical rejection rate	12.90%	24.47%
Total rejection rate	**27.36%**	**27.85%**
System utilization	**87.13%**	**84.33%**

Table 3.3: Performance measures computed from year 2000 data and the simulation model when the capacity of the SICU is zero

2000, only 40 records (about 3%) are incomplete and and two patients have irregular LOS ($>$ 2 years). After cleaning these data points, the resulting utilization rate is computed to be 87.13%. We expect that this number represents a very close estimate for the true utilization rate since only a few patient records are excluded from our calculation.

Table 3.3 provides the results from the ICU data in 2000 and from the simulation when the capacity of the SICU is set to be zero. As can be seen, our simulation model is able to match the total and seasonal rejection rates of *all* patients as well as the utilization rate from the actual data. However, the rejection rates of surgical and medical patients are different. In particular, the medical rejection rates obtained from the data are lower than those computed from the simulation. This is because we treat the admission processes of both types of patients indifferently by setting the capacity of the SICU to zero, while the actual ICU does give priority in admissions to medical patients.

To capture this service-based admission policy, we increase the capacity of the SICU in the simulation model. We eventually found that, with the capacity of four beds provided to the SICU, almost all the results from our simulation becomes consistent with those from the actual data. This is illustrated in Table 3.4. The exceptions are the winter and non-winter medical rejection rates, which are relatively different

from the actual results. We will discuss this discrepancy shortly. Also observe that the adjustment in the SICU capacity has a small effect on the total rejection rates and the utilization level, which should be attributed to the difference in the LOS distributions between surgical and medical patients.

Parameter	Result from year 2000	Simulation Result
Winter surgical rejection rate	41.98%	42.68%
Winter medical rejection rate	11.58%	15.82%
Total winter rejection rate	28.55%	30.50%
Non-winter surgical rejection rate	33.45%	31.94%
Non-winter medical rejection rate	13.86%	10.64%
Total non-winter rejection rate	26.69%	24.62%
Surgical rejection rate	36.21%	35.86%
Medical rejection rate	12.90%	13.12%
Total rejection rate	27.36%	27.03%
System utilization	87.13%	83.79%

Table 3.4: Performance measures computed from year 2000 data and the simulation model when the capacity of the SICU is four

To see if the SICU capacity of four beds is realistic, we now investigate the actual number of surgical patients that were cared outside the ICU in 2000. Fig.3-7 shows the number of off-service ICU surgical patients that stayed in other medical units with respect to the time in year 2000. According to this data, the average utilization of the off-service beds is 1.34 bed, and the SD of the occupancy outside the ICU is equal to 1.52 bed. Clearly, the figure implies that the number of beds allowed for surgical diversions from the ICU tends to be time-varying. Since in simulation we fix the capacity in the SICU throughout, this could be the reason why the seasonal rejection rates from simulation, especially those of medical patients, deviate from the seasonal rejection rates obtained from the actual data. Nevertheless, the figure suggests that the SICU capacity of four beds could be reasonably used as an approximate average capacity allowed by medical units outside the ICU at CHB in 2000. This as well shows the validity of our model since fixing the realistic average SICU capacity allows the simulation to give the accurate estimates for almost all rejection rates and system utilization compared to the actual results.

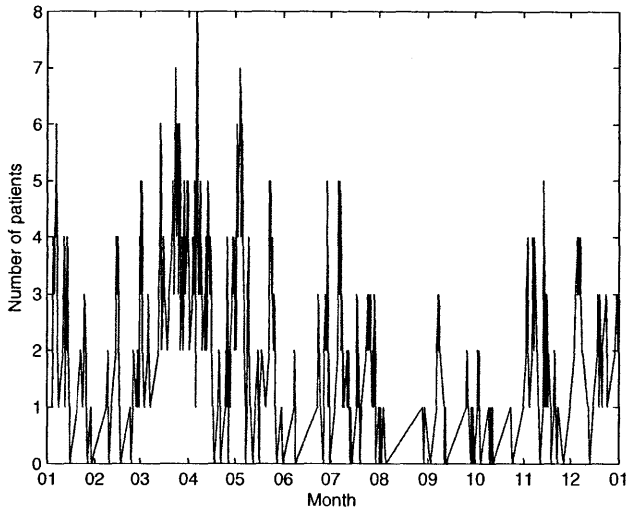


Figure 3-7: The number of off-service ICU surgical patients in 2000

To summarize, we have verified that our simulation model can be calibrated to reliably analyze the ICU system at CHB. In the next chapter, we will discuss using this simulation model to evaluate the performance of the cap-based scheduling policies.

Chapter 4

Performance of the Cap-Based Admission Control Policies

In this chapter, we discuss the performance of the two cap-based policies which we introduced in Chapter 2. The simulation model developed in the previous chapter is used to investigate various performance measures of the policies. Our major finding is that the cap-based scheduling policies are able to improve the rate of rejections in the ICU, at the cost of increasing mean waiting time of scheduled patients. We also find that the efficiency of both policies depends on the level of system utilization.

Next, we formally introduce the two cap-based policies and describe cap allocation criteria as well as the potential impacts of the policies on the flow of patients in the ICU.

4.1 Uniform Cap Policy

The uniform cap policy (UCP) intends to reduce variability in surgical demand by placing a limit on the *total* number of surgical cases that can be scheduled to the ICU on a single day. Although this policy has been implemented at CHB since 2003, no attempt has been made to track its impacts on the unit. We will use our simulation model in investigating the performance of the UCP later on in this chapter.

4.1.1 Notation

Let a uniform cap (UC) be a vector of five integer values, where the n^{th} element corresponds to the number of scheduled surgical admissions allowed on the n^{th} weekday (the weekdays start on Monday and end on Friday). The term "cap" is used when referring to the number of scheduled cases allowed in a day. For example, a UC = [5 5 5 5 4] means that the cap allows five cases from Monday through Thursday and four cases on Friday. There is no cap on Saturdays and Sundays since the ORs are not open for scheduled surgeries over the weekends.

4.1.2 Cap Allocation Rules

The only rule required for allocating a UC is that it must be higher than the weekly arrival rate of scheduled patients to ensure the stability of the system. In this context, system stability means that the size of the queue of scheduled patients never grows unbounded. Caps become redundant once they are raised so high that they exceed the maximum number of preplanned surgeries that are allowed to enter the ICU per day (which is clearly upper-bounded by the number of ORs).

Note that there can be several combinations of caps that satisfy this stability condition. Nevertheless, we choose to evenly spread caps throughout the week in order to smoothen the demand from scheduled patients for an ICU. Besides, an evenly-distributed cap allows patients to be scheduled on a consistent basis, which could save the amount of waiting time better than a cap that simply releases one or two huge batches of patients into the ICU over a week. In fact, this method of cap allocation is used by the ICU at CHB.

4.1.3 Expected Impacts of the Policy

The initial impact we expect from the UCP upon the ICU is the smoother demand pattern of scheduled surgeries who requires post-operative ICU beds. This would allow the policy to be able to reduce the number of days on which a very high number of rejections takes place, which is caused by the huge swing in the number

of daily arrivals to the ICU. As a result, we believe that the rejection rate in the ICU with the UCP would be lower than that in the ICU without caps. Meanwhile, using caps to restrict the number of daily admissions could result in a larger queue of scheduled surgical patients, which in turn increases their mean waiting time[1]. In fact, the closer the caps are to the average weekly arrival rate of scheduled patients, the longer the waiting time those patients have to wait before their surgery can start. It is important to understand this trade-off and we will use simulation method for this goal.

4.2 Service-Specific Cap Policy

Aiming to reduce variability in demand for an ICU, the UCP does not use information about the LOS of surgical patients in planning caps. However, the statistical analysis of the LOS in Section 2.8.3 of Chapter 2 indicates that the LOS of surgical patients are heterogeneous with respect to services. It is known that patients with long LOS, although representing a small portion of all patients in the unit, occupy ICU resources much more than those with short LOS (Ryckman et al. [32] and Stricker et al. [33]). The admission of these long-stay patients, when uncontrolled, would contribute to the likelihood of system overcrowding, which consequently leads to rejections in the unit.

Therefore, we propose the service-specific cap policy (SSCP) as an extension of the UCP that aims to control the scheduling of different groups of surgical patients, which are classified based on their LOS statistics. The goal of the SSCP is to limit the number of long-stay surgical cases that can be scheduled to a single day.

Now, we introduce the formal cap notation as well as discuss the allocation rules for the SSCP.

[1]This claim follows directly from Little's Law, which states that the steady-state average queue length is equal to the arrival rate times the steady-state average waiting time.

4.2.1 Notation

Throughout this chapter, a service-specific cap (SSC) consists of four cap vectors. Each vector contains five integer values, where the n^{th} element corresponds to the cap on the n^{th} weekday. Like a UC, no SSC is administered over the weekends. There are three groups of surgical patients as classified based on their average LOS in Section 2.8.3 of Chapter 2. Out of the four vectors, the first three belong to the respective groups of surgical patients, and the last one is the total cap vector that controls the total number of scheduled cases allowed on each day in the week. A total cap vector offers flexibility in scheduling by the SSCP and will be in effect only if its capacity is less than the sum of the capacity from the first three cap vectors. One can think of a total cap vector as a UC vector embedded in a SSC. The four cap vectors of a SSC can be expressed as a matrix of 4×5 dimensions. An example of a SSC is

$$\begin{pmatrix} 3 & 3 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 \\ 7 & 6 & 6 & 6 & 6 \end{pmatrix},$$

where the first three rows represent the cap for Group 1, 2, and 3 of surgical patients respectively and the last row represents the total cap. Note that these numbers were chosen arbitrarily just to provide an example of the SSC. Certain elementary rules for choosing a SSC are provided in the following section.

4.2.2 Cap Allocation Rules

Similar to the case of a UC, the cap vectors of each surgical group must be planned so that they are higher than the group's weekly arrival rate in order to ensure the stability condition. A cap vector of a SSC becomes redundant when it is set higher than the capacity of the ORs provided. It is also redundant when the sum of the cap vector of each surgical group is larger than the caps allowed by the total cap vector.

There are several candidates of caps that satisfy this condition. However, like the

case of the UCP, we always spread a SSC through each day of the week to smoothen the demand of surgical patients and to not unnecessarily prolong their waiting times.

4.3 Computational Results and Discussion

4.3.1 Impacts of the Cap-Based Policies

The simulation model is used to investigate the performance of the ICU under the two cap-based policies. The simulation scenario in this section is built on the ICU environment and data in 2000. In particular, the main ICU capacity is set to be 16, and the capacity provided to the SICU is equal to four beds. To guarantee the stability condition, a UC is set to be [5 5 5 5 5]. This choice of the UC is fixed through both the winter and non-winter seasons in order to imitate the caps used in the ICU, which are not adapted seasonally. A SSC is also fixed through seasons with the total cap vector equal to the UC vector. Indeed, the SSC is set to be

$$\begin{pmatrix} 3 & 3 & 2 & 2 & 2 \\ 1 & 1 & 2 & 1 & 2 \\ 1 & 1 & 1 & 2 & 1 \\ 5 & 5 & 5 & 5 & 5 \end{pmatrix}.$$

Each cap vector is chosen to meet the stability condition and, at the same time, to not exceed the capacity of the ORs provided from the block time. Note that at this stage we do not attempt to optimize the rejection rates and just look at one particular SSC.

100 independent sample paths of arrivals to the ICU, each with the time horizon of two years, are generated according to the data in 2000. The sample paths are simulated with the no-cap policy, UCP, and SSCP. The results from each policy are summarized in Tables 4.1 to 4.4.

Most of the results shown in these tables are consistent with our expectation on the impacts of the cap-based policies. In particular, Table 4.4 and 4.1 show that both

Policy	Mean of the total rejection rate	SD of the total rejection rate	95% confidence interval
No-cap	26.95%	2.65%	[26.43%, 27.47%]
UCP	25.82%	2.55%	[25.32%, 26.32%]
SSCP	25.46%	2.81%	[24.92%, 26.01%]

Table 4.1: Total rejection rate statistics obtained from different scheduling policies

Policy	Surgical rejection rate	Medical rejection rate	Total rejection rate
No-cap	35.73%	13.12%	26.95%
UCP	33.65%	13.49%	25.82%
SSCP	33.19%	13.29%	25.46%

Table 4.2: Total rejection rate statistics of surgical and medical patients obtained from different scheduling policies

Group	Mean waiting time (days)		
	No cap	UCP	SSCP
1[a]	1.33	2.90	9.45
2[b]	1.13	2.71	4.62
3[c]	0.96	2.52	7.52

[a] Short LOS patients: neurosurgical, ORL, plastics surgery, urology, and OMFS

[b] Intermediate LOS patients: orthopedic, trauma, and IntRadio

[c] Long and highly variable LOS patients: general surgery and other surgery

Table 4.3: Mean waiting times of surgical patients obtained from different scheduling policies

Policy	SD of elective surgery demand	SD of unit occupancy	Utilization	CV of unit occupancy	Percentage of saturation period
No-cap	1.96	2.360	83.79%	0.451	22.55%
UCP	0.86	2.273	84.81%	0.429	23.75%
SSCP	0.71	2.249	84.88%	0.424	23.38%

Table 4.4: Other performance measures obtained from different scheduling policies

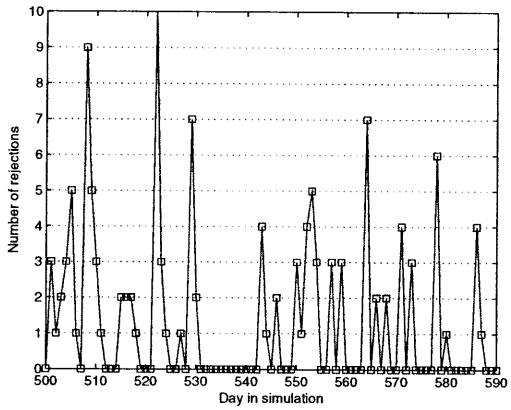
policies are able to reduce the variability in elective surgery demand and, consequently, the total rejection rate in the ICU. This is, however, achieved at the expense of the longer mean waiting times of scheduled surgical patients as evidenced in Table 4.3.

Specifically, the SSCP gives the lower rejection rate as well as the longer mean waiting times between the two policies. Notice that the utilization rates in the ICU with caps are slightly higher than the ICU without caps, which is probably because caps prevent rejections in the ICU and allow the unit to utilize its resources better.

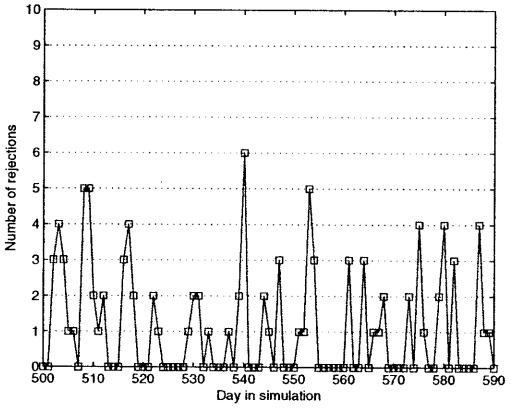
As can be seen in Table 4.2, both cap-based policies are able to reduce the surgical rejection rate. This is an expected result since the implementation of caps evenly spreads the arrivals of surgical patients over time and as a result increases their acceptance rate to the ICU. In particular, the SSCP performs slightly better than the UCP in this regard. However, the medical rejection rates are increased after using caps in the ICU. This is an interesting phenomena and we will investigate it further in Section 4.3.3 when the utilization rate in the ICU is varied.

As the number of arrivals in 2000 is divided into 1227 surgical arrivals and 744 medical arrivals, the UCP and SSCP would be able to cut down the number of surgical patients that could have been diverted from 438 to 413 and 407 patients, respectively. Meanwhile, the two policies slightly increase the number of medical rejections from 98 to 100 and 99 cases. Overall, the UCP and the SSCP can reduce the total number of rejections/diversions from 536 to 513 and 506, which are amount to 32 and 30 fewer annual rejections/diversions on average, respectively. Observe that the overall rejected/diverted patients in the ICU with caps are still high. This implies that the ICU at CHB would have encountered a large number of rejections/diversions despite the implementation of the UCP in 2003, which probably led to the unit expansion in 2005.

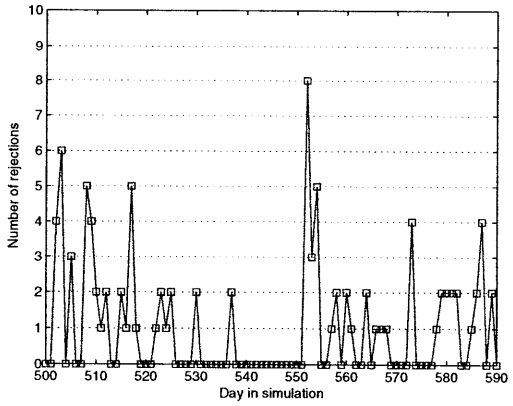
Moreover, Table 4.4 indicates that the implementation of caps reduces the CV of the unit occupancy (namely the SD of unit occupancy divided by the mean of unit occupancy), which implies that the policies offer a better control to the utilization of ICU resources. In addition, we notice from the table that the percentage of time for which the ICU is full increases after using caps. This last result could be attributed to the fact that caps enable the ICU to serve more patients, which possibly leads to more saturation times in the system. However, as we will see in Section 4.3.3, the implementation of caps is indeed capable of decreasing the full-occupancy period



(a) No cap



(b) UCP



(c) SSCP

Figure 4-1: Number of rejected patients from one sample path under different scheduling policies

when the utilization level of the ICU is not too high.

Fig.4-1 shows that applying caps can reduce the number of days with high rejections in the ICU(say, more than five cases). In other words, smoother demand prevents the possibility that the huge batches of arrivals would enter the ICU on a single day, which consequently decreases the number of rejections that might take place during overcrowded hours. This contribution of caps could be the main reason why the total rejection rates are dropped despite the increase in saturation periods after implementing caps in the simulated system.

Remarks on waiting times of scheduled patients

As discussed in Section 3.2.3 of Chapter 3, the mean waiting time of scheduled patients in our simulation model is likely to be shorter than it should be in the real ICU system. This is because the simulation assumes that each elective surgical patient is always scheduled to the very first open OR (if caps permitted) regardless of surgeons' schedules. Meanwhile, the scheduling process in reality takes into account the choice of surgeons that elective patients might request for their surgery (which is always the case), and then schedules them according to the block of those selected surgeons. In addition, the most recent OR might not be available for scheduling a case that requires an ICU since it could have been completely filled by non-ICU cases. These reasons explain why scheduled surgical patients in the simulation are likely to wait shorter than those in the real system. Nevertheless, we expect that real waiting times will increase after implementing the UCP and the SSCP as the simulated ones do.

4.3.2 Impacts of Varying Caps

We now investigate the performance of the cap-based polices at different cap levels in the ICU when arrival rates are fixed. The 100 sample paths generated in the previous section are used to simulate the ICU with the UCP and SSCP. Table 4.5 summarizes the total rejection rates and the SD of elective surgery demand, and the mean waiting times of each group of surgical patients are presented in Table 4.6. The detail of each

SSC used in this experiment is as follows:

$$\begin{pmatrix} 3 & 3 & 2 & 2 & 2 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 5 & 5 & 5 & 5 & 4 \end{pmatrix}, \begin{pmatrix} 3 & 3 & 2 & 2 & 2 \\ 1 & 1 & 2 & 1 & 2 \\ 1 & 1 & 1 & 2 & 1 \\ 5 & 5 & 5 & 5 & 5 \end{pmatrix}, \begin{pmatrix} 3 & 3 & 2 & 3 & 2 \\ 1 & 1 & 2 & 1 & 2 \\ 2 & 1 & 1 & 1 & 1 \\ 6 & 5 & 5 & 5 & 5 \end{pmatrix},$$

$$\begin{pmatrix} 3 & 2 & 3 & 2 & 3 \\ 1 & 2 & 1 & 2 & 1 \\ 2 & 2 & 1 & 1 & 1 \\ 6 & 6 & 5 & 5 & 5 \end{pmatrix}, \begin{pmatrix} 3 & 2 & 3 & 3 & 3 \\ 1 & 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 & 1 \\ 6 & 6 & 6 & 5 & 5 \end{pmatrix}.$$

Cap	Total rejection rates (%)		SD of scheduled demand	
	UCP	SSCP	UCP	SSCP
[5 5 5 5 4]	25.75	25.42	0.74	0.67
[5 5 5 5 5]	25.82	25.46	0.86	0.71
[6 5 5 5 5]	25.92	25.51	1.07	0.93
[6 6 5 5 5]	26.15	25.75	1.19	0.99
[6 6 6 5 5]	26.27	25.85	1.29	1.11
No cap	26.95		1.96	

Table 4.5: Total rejection rates in the ICU with the cap-based polices at various cap levels. The cap vectors shown in the first column correspond to the UC and the total cap vectors in the SSC.

Cap	Mean waiting time (days)					
	Group 1		Group 2		Group 3	
	UCP	SSCP	UCP	SSCP	UCP	SSCP
[5 5 5 5 4]	4.61	11.18	4.46	6.48	4.28	9.07
[5 5 5 5 5]	2.69	9.06	2.51	4.53	2.35	7.42
[6 5 5 5 5]	2.14	3.68	1.95	4.45	1.80	7.56
[6 6 5 5 5]	1.91	3.22	1.70	4.31	1.56	2.85
[6 6 6 5 5]	1.74	2.21	1.52	4.70	1.40	2.85
No-cap	1.32		1.06		0.93	

Table 4.6: Mean waiting times of surgical patients in the ICU with the cap-based policies at different cap levels. The cap vectors in the first column correspond to the UC and the total cap vectors in the SSC.

The tables indicate that, as the size of caps grows, the demand of scheduled surgeries becomes more variable and the rejection rates increase, while the mean waiting times tend to decrease. The latter is an intuitive result since larger caps allow more patients to enter the ICU on a single day, and this in turn leads to a shorter mean waiting time. Meanwhile, raising caps provides more flexibility in scheduling patients, which incurs additive variability to the demand of scheduled surgeries and therefore amplifies the rate of rejections in the ICU. The trade-off between the rejection rate and the mean waiting time must be addressed when deciding which size of caps should be used in scheduling ICU surgical patients.

4.3.3 Performance of the Cap-Based Policies as a Function of System Utilization

The times during which the unit capacity reaches its peak is undesirable as any arrival in this period will be blocked from entering the unit. We speculate that the ICU is likely to face a saturation period more often with highly-variable demand, since it could generate a large number of ICU entries on a single day that would fill in all space in the ICU at once. By using caps in the ICU, our results in Section 4.3.1 show that they are able to spread the ICU demand evenly over a period and thus feed the unit with move even numbers of daily arrivals. One might conjecture that a smoother demand as a result of implementing caps would be capable of lowering the chance that the ICU would become full.

However, as seen in Table 4.4 from Section 4.3.1, the percentage of congestion periods in the ICU increase after applying caps. We believe this is because the ICU system that we considered in that section is highly utilized ($\sim$82%), so the amount of the baseline demand could be so large that an evenly-distributed pattern of the demand itself might not be able to help decreasing saturation times. Instead, implementing caps in such a heavily-utilized ICU could increase the chance that the system becomes overcrowded, since spreading the massive demand from many *heavily-loaded* days possibly raises the resource utilization on the nearby *lightly-loaded* days

to the capacity limit. For these reasons, we expect that caps might not be as effective when the ICU is heavily utilized as in the case where the system is operating at a relatively lower utilization level.

Nevertheless, it is important not to interpret any increasing chance of system congestion as a disadvantage from using caps. Indeed, we view this as an improvement in system utilization, as caps assist an ICU in better utilizing its resources through the increase in the admission rate. The chance that the unit might become overcrowded more frequently after applying caps is probably due to the increase in the utilization rate of an already highly-utilized system.

We examine via simulation the performance of the cap-based policies at different utilization levels. The sample paths of arrivals are drawn from the data in year 2000, but with varied arrival rates in order to generate various degrees of system utilization. For each arrival rate, a set of 100 sample paths are simulated under the no-cap policy, UCP, and SSCP. All caps are set to be the tightest yet satisfy the stability condition.

Utilization level	Total rejection rate (%)		
of the no-cap policy (%)	No cap	UCP	SSCP
89.76	44.34	43.80	43.72
87.54	36.65	35.82	35.78
86.00	32.06	30.74	30.62
83.79	26.95	25.75	25.42
80.90	20.80	19.15	18.95
77.57	16.05	14.51	14.20
74.77	12.68	11.01	10.69
72.41	10.29	8.24	8.08
69.13	7.45	5.75	5.47
65.15	5.32	3.92	3.68
60.44	3.27	2.14	2.12
56.12	1.75	1.06	1.04

Table 4.7: Total rejection rates from different scheduling policies with respect to the varying levels of system utilization

Table 4.7 summarizes the total rejection rates at various utilization levels obtained from the simulation. The decrease in the total rejection rates after applying caps as a function of utilization is given in Fig.4-2. As can be seen from the figure,

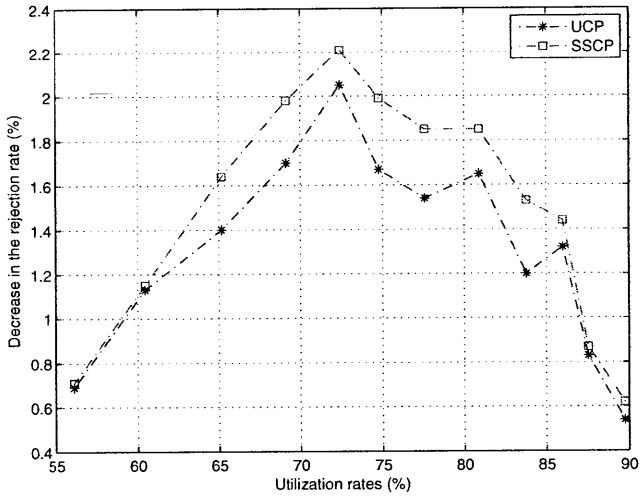


Figure 4-2: Decrease in the rejection rates after implementing the UCP and SSCP at different utilization levels

both cap-based policies are most effective in the medium utilization regime ($\sim$ 70-75%), as opposed to their performance in the high utilization regime ($>$ 75%). The improvements become smaller in the low utilization regime ($<$ 70%) since the rejection rate in the ICU without caps is already low, so there is not much space to improve after applying the policies.

Utilization level	SD of the unit occupancy			CV of the unit occupancy		
of the no-cap policy (%)	No cap	UCP	SSCP	No cap	UCP	SSCP
89.76	1.887	1.862	1.859	0.336	0.330	0.330
87.54	2.079	2.039	2.037	0.383	0.370	0.369
86.00	2.205	2.118	2.125	0.410	0.390	0.391
83.79	2.360	2.282	2.260	0.451	0.431	0.426
80.90	2.520	2.415	2.409	0.498	0.470	0.469
77.57	2.685	2.585	2.582	0.554	0.525	0.524
74.77	2.790	2.689	2.676	0.596	0.567	0.564
72.41	2.845	2.738	2.730	0.620	0.596	0.596
69.13	2.918	2.799	2.781	0.675	0.637	0.634
65.15	2.988	2.874	2.829	0.734	0.697	0.685
60.44	3.018	2.891	2.874	0.799	0.762	0.757
56.12	2.953	2.790	2.769	0.842	0.792	0.787

Table 4.8: SD and CV of the unit occupancy from different scheduling policies with respect to the varying levels of system utilization

Utilization level	Percentage of unit saturation periods		
of the no-cap policy (%)	No cap	UCP	SSCP
89.76	36.95	38.24	38.30
87.54	30.61	32.32	32.26
86.00	27.02	28.72	28.41
83.79	22.55	23.74	23.44
80.90	17.62	18.56	18.45
77.57	13.41	13.79	13.62
74.77	10.67	10.48	10.37
72.41	8.42	8.00	7.79
69.13	6.19	5.82	5.62
65.15	4.30	3.85	3.60
60.44	2.64	2.07	2.06
56.12	1.43	1.03	0.99

Table 4.9: Percentage of the saturation period from different scheduling policies with respect to the varying levels of system utilization

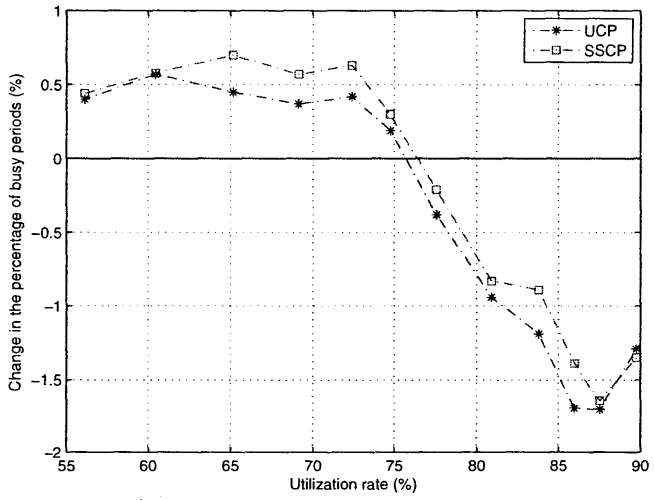


Figure 4-3: Change in the percentage of the saturation period in the ICU with different scheduling policies as a function of utilization

We now proceed to analyze the CV of the unit occupancy at different system utilization levels. As can be seen in Table 4.8, caps reduce the CV of the unit occupancy in the ICU consistently at all levels of system utilization. This implies that the ICU can achieve a better control on its unit occupancy level by using caps. Notice that the CV tends to increase as the utilization rate becomes smaller. This is

partly because the CV itself is inversely proportional to utilization levels. Another reason is that the SD of the unit occupancy tends to increase when the utilization goes down since the number of busy beds can vary from low to high over a period of time, as opposed to the case when the system is highly-utilized, which forces the corresponding unit occupancy to often concentrate around the high values.

Table 4.9 shows the percentage of the saturation period in the ICU under three different scheduling policies with respect to varying utilization levels. The change in this performance measure is illustrated in Fig.4-3. It is clear from the figure that the implementation of the cap-based policies in a highly-utilized system can result in the extended period of saturation. As the utilization rate descends pass 75%, both policies begin to reduce the full-occupancy period. In addition, our results indicate that the SSCP consistently performs better between the two policies as far as the improvement in the probability of ICU overcrowding is concerned.

Utilization level	Surgical rejection rates (%)			Medical rejection rates (%)		
of the no-cap policy (%)	No cap	UCP	SSCP	No cap	UCP	SSCP
89.76	54.84	52.97	52.78	27.78	29.36	29.43
87.54	46.44	44.30	44.35	21.24	22.49	22.29
86.00	41.64	38.77	38.68	17.01	18.11	17.99
83.79	35.73	33.51	33.15	13.12	13.54	13.27
80.90	28.61	25.95	25.58	8.40	8.37	8.44
77.57	22.72	20.25	19.83	5.53	5.45	5.31
74.77	18.37	15.85	15.39	3.73	3.42	3.31
72.41	14.89	12.36	11.25	2.52	2.05	1.80
69.13	11.27	8.74	8.30	1.44	1.06	1.04
65.15	8.23	6.05	5.71	0.74	0.56	0.48
60.44	5.08	3.33	3.33	0.41	0.28	0.23
56.12	2.77	1.68	1.65	0.13	0.08	0.06

Table 4.10: Rejection rates of surgical and medical patients from different scheduling policies with respect to the varying levels of system utilization.

We now investigate the rejection rates of surgical and medical patients from different scheduling policies at various utilization levels, both of which are presented in Table 4.10. The decrease in the rejection rates of both types of patients with respect to system utilization are given in Fig.4-4 and Fig.4-5, respectively. As evidenced in

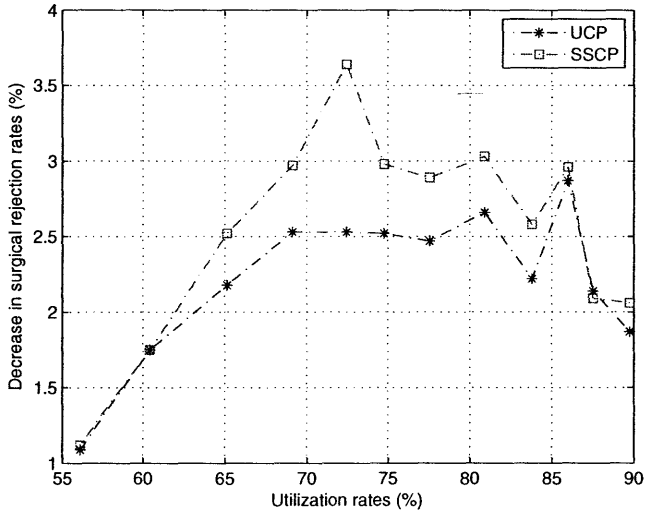


Figure 4-4: Decrease in the surgical rejection rates after implementing the UCP and SSCP at different utilization levels

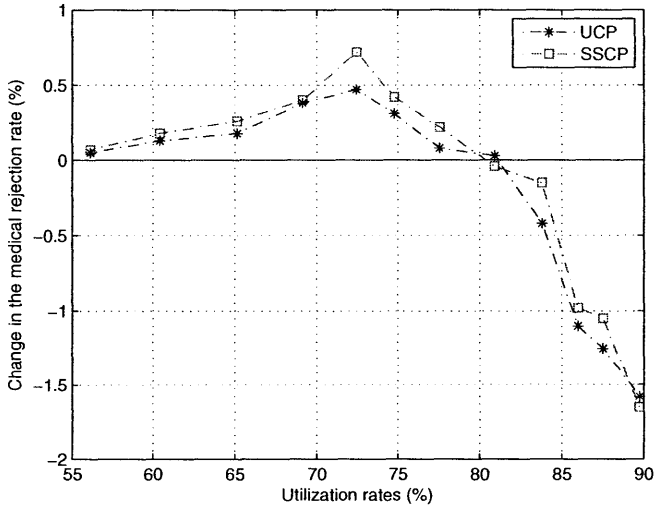


Figure 4-5: Change in the medical rejection rates after implementing the UCP and SSCP at different utilization levels

Fig.4-4, both cap-based policies reduce the surgical rejection rate at all utilization levels. Similar to our discussion in Section 4.3.1, we believe that this is because caps evenly distribute the arrivals of surgical patients to the ICU over time, which consequently increases their rate of admission. On the other hand, Fig.4-5 shows that caps

do not always lead to the decrease in medical rejection rates. In particular, the figure suggests that using caps in the ICU when system utilization is high ($> 80\%$) results in the increase in medical rejection rates, whereas the rates become smaller when using caps in the relatively lower utilization regime. This is indeed a very interesting result; however, we do not have a full answer for it yet. One plausible explanation is that applying caps to an already highly-utilized system allows more surgical patients to utilize ICU resources, which would raise the unit occupancy to the peak more often over time. Since medical patients randomly arrive to the ICU, it is more likely that they will be blocked from entering the unit as a result of more overcrowding times. On the other hand, applying caps to the ICU when system utilization is not too high can help reducing overcrowding periods, according to the result in Fig.4-3. Therefore, medical arrivals tend to see a vacant bed more often when they arrive to the ICU in this case. Note that the rejection rate of medical patients is decreased the most when the system utilization is about $70\% - 75\%$.

Our results in this section substantiate why the cap-based policies are not as efficient in a heavily-utilized ICU as they are in a system with moderate load. As such, an ICU should not expect a good performance from the policies while still over-utilizing the available resources. Indeed, the ICU administration should consider increasing the capacity to match the demand for critical care as the first-line attempt to improve the flow of patients. This way would prevent the overuse of ICU resources and thereby allow more significant impacts from caps. In addition, one needs to be careful when using caps in a highly-utilized system since they could cause more rejections for medical patients.

4.4 Impacts of Reducing the LOS

The departure of ICU patients can be delayed for a variety of reasons. For example, patients who are ready to leave in the middle of the night usually continue to stay further and leave in the morning of the next day since the ICU usually prefers to discharge their patients during the daytime. In addition, patients might need to stay

in the ICU pass their projected departure times until floor units are able to find beds to host them in the post-ICU period. As a result, the LOS obtained from the ICU census are likely to overestimate the minimum required staying times of patients.

Distribution of T'	Total rejection rate (%)		
	No cap	UCP	SSCP
0	26.95	25.82	25.46
U(0, 4)	25.32	24.53	24.18
U(0, 8)	24.32	23.18	22.71
U(0, 12)	23.02	21.91	21.56
U(0, 16)	21.63	20.55	19.95
U(0, 20)	20.03	18.92	18.61
U(0, 24)	18.62	17.36	17.23

Table 4.11: Total rejection rates from different scheduling policies with respect to the distribution of T'

We are interested in evaluating the performance of the ICU system in which the LOS are reduced by a certain amount of time. The same set of 100 sample paths that were generated in Section 4.3.1 are used to simulate the ICU with no cap, UCP, and SSCP. The only difference is that each patient in a sample path is now assumed to leave the ICU earlier by a random amount of time T', which is uniformly distributed between 0 and a hours, i.e., $T' \sim \mathrm{U}(0, a)$. The new LOS is therefore equal to $\max\{0, \mathrm{LOS} - T'\}$. The specifications of the UC and SSC are also set to be the same as those used in Section 4.3.1.

Table 4.11 shows the rejection rates in the ICU under three scheduling policies according to the varying values of a. The drop in rejection rates after uniformly reducing the LOS implies that the ICU can benefit from discharging their patients earlier. In fact, the results indicate that the rejection rates are already decreased if the ICU is able to release their patients on average a few hours earlier. Moreover, since the LOS is artificially prolonged by floor bottlenecks, this analysis defines the limit of performance improvements aimed at removing ICU outflow obstructions.

Chapter 5

Queueing Model of the ICU at Children's Hospital Boston

In this chapter, we develop a discrete-time queueing model that can be used to analyze the ICU system at Children's Hospital Boston (CHB). The main purpose of formulating the model is to compute performance measures, such as the rejection rates and the mean waiting time, and compare them with the simulation results. We will show a strong agreement of both approaches. As a result, a queueing model can be a useful alternative for the simulation-based model.

5.1 Conceptual Framework

5.1.1 Outline of the Queueing Model

The queueing system consists of two modules: the queue of surgical (scheduled) patients and the ICU. Time evolution in the system is discretized into days. At the beginning of the day, the queue sends a group of surgical patients to the ICU. Then, a batch of surgical patients, which is distributed as a Poisson random variable, arrives to the queue. New arrivals are rejected if the queue already reaches its maximum capacity. Surgical patients admitted to the queue wait until they are sent to the ICU according to the First-In-First-Out (FIFO) discipline.

At the ICU, a day starts off as the ICU receives arrivals of surgical patients who are sent from the queue on the same day. Subsequently, a batch of medical patients, which is also distributed as a Poisson random variable, arrives to the ICU. New patients are admitted to the ICU unless they arrive to a full unit. The length of stay (LOS) of each admitted patients is geometrically distributed. At the end of the day, the ICU checks all current patients to see if any can be discharged from the unit.

5.1.2 Modeling Assumptions

Several assumptions are made so that the queueing model becomes computationally tractable. First, we make no distinctions among different services of surgical arrivals since otherwise additional state variables of the queue length associated with each surgical service are needed. Similarly, there is no seasonality in the queueing model so as to limit the complexity of time evolution. We also assume the same LOS distribution for both surgical and medical patients in order to eliminate the need of an additional state that tracks the number of both patients in the system. In addition, we assume that the number of arrivals in each day is independent and identically distributed (i.i.d.) and that the LOS of each patient is geometrically distributed so that we can establish the Markov property in the queueing system.

5.1.3 Solution Methods

We will show that the queue length and bed occupying processes form a Markov chain. This property allows us to numerically compute the steady-state probabilities of both chains. The stationary probability of the queue length process is used to compute the mean queue length and the mean waiting time in the steady state, while the stationary probability of the bed occupying process is used to calculate the mean rejection rate in the steady state.

5.1.4 Limitations of the Queueing Model

Although the queueing model is developed to analyze the ICU at CHB, there are several aspects of the actual ICU that are not addressed by the model. They are summarized as follows.

1. Since our queueing model is discrete-time, patients arrive to and depart from the ICU as batches in each time step rather than continuous-time processes, which could serve as a better representation of arrivals and departures in the real system.

2. The model does not consider the seasonality of the arrivals and the LOS, which in fact takes place in the ICU at CHB according to the statistical analysis in Section 2.7.2 and Section 2.8.2 of Chapter 2.

3. The model does not distinguish the services required by surgical patients, while the actual ICU definitely does.

4. The model treats surgical and medical patients indifferently regarding their LOS distributions, while there is a clear difference in the LOS among different types of patients as suggested by the LOS data analysis in Section 2.8.3 of Chapter 2.

5. The model assumes that the LOS distributions are memoryless, but the analysis of the tail distributions of the LOS in Section 2.8.4 of Chapter 2 does not imply this property.

Despite these simplifications, we shall see in the results and discussion section that the queueing model provides estimates for performance measures that are consistent with the results from simulation. Moreover, the model is capable of capturing the trade-off between rejection rates and mean waiting times by the uniform cap policy (UCP) as suggested by the simulation in the previous chapter.

5.2 Dynamics of the System

5.2.1 Notation

C_{queue}	Maximum capacity of the queue of surgical patients
C_{ICU}	Number of beds in the ICU
D_t	Day of the week at time t
$\text{cap}(D_t)$	Number of cases (cap) allowed on day D_t
A_t^1	Number of surgical arrivals at time t
A_t^2	Number of medical arrivals at time t
S_t	Number of surgical patients sent from the queue to the ICU at time t
Q_t	Queue length at time t
I_t	Number of busy beds at time t

5.2.2 Queue of Surgical Patients

One time step of our discrete-time queueing model corresponds to one day and is indexed by a discrete index $t \in \mathbb{Z}^+$. Let $D_t \in \{1, 2, \ldots, 7\}$ be a day of the week at day t where $D_t = t \bmod (7) + 1$. At the beginning of day D_t, there are S_t surgical patients leaving from the queue to enter the ICU. The maximum number of patients from the queue that are allowed to enter the ICU on day D_t is referred to as the cap on D_t, denoted by $\text{cap}(D_t)$. With the cap being administrated, S_t^1 will be equal to the minimum of the number of patients in the queue and $\text{cap}(D_t)$, i.e.,

$$S_t = \min\left(Q_t, \text{cap}(D_t)\right). \tag{5.1}$$

After sending S_t patients to the ICU, there will be A_t^1 arrivals of surgical patients entering the queue during day t. We assume that the queue discipline is First-In-First-Out (FIFO) and that the queue is truncated to the capacity of C_{queue}. Thus, patients who arrive and see a full queue are rejected. Define Q_t to be the queue length at the end of day t. The actual number of arrivals to the queue on day t is equal to

$$\min\{A_t^1, C_{queue} - Q_t + S_t\}.$$

As a result, the queue length process $\{Q_t | t \in \mathbb{Z}^+\}$ can be described by the following equation:

$$Q_{t+1} = Q_t - S_t + \min\{A_t^1, C_{queue} - Q_t + S_t\}. \tag{5.2}$$

5.2.3 ICU

The time evolution in the analysis of the ICU is the same as that in the queue. At the beginning of day t, the queue sends out S_t patients to the ICU. Each surgical patient is admitted if there is an empty bed available in the unit. Let C_{ICU} be the capacity of the ICU and let I_t be the number of busy beds in the ICU at the end of day t. The number of surgical admissions on day t is equal to

$$\min\{S_t, C_{ICU} - I_t\}.$$

In addition to surgical patients, there are A_t^2 medical patients that come to the ICU at time t. Similar to surgical patients, a medical patient is admitted if there is space available in the unit. We assume that the admission of medical patients are done after surgical patients' in order to give priority to scheduled patients. If the unit is filled completely by surgical patients, no medical arrivals will be admitted. The number of medical admissions on day t is

$$\min\Big\{A_t^2, C_{ICU} - I_t - \min\{S_t, C_{ICU} - I_t\}\Big\}.$$

At the end of day t, a random number of current patients, including the newly-admitted surgical and medical patients on that day, will leave the ICU. We denote this quantity by B_t. As a result, the bed occupancy process $\{I_t | t \in \mathbb{Z}^+\}$ evolves according to

$$\begin{aligned} I_{t+1} &= I_t + \min\{S_t, C_{ICU} - I_t\} \\ &\quad + \min\Big\{A_t^2, C_{ICU} - I_t - \min\{S_t, C_{ICU} - I_t\}\Big\} - B_t. \end{aligned} \tag{5.3}$$

5.3 Markov Chain Model of the Queueing System: Stochastic Primitives and State Transitions

We assume that the number of arrivals in each day is distributed as a Poisson random variable with mean λ_1 for surgical patients and λ_2 for medical patients. Moreover, we assume that the LOS of each patients is i.i.d. and geometrically distributed with mean $1/\mu$. Now, let us provide the analysis of the transition probability of the queue length and the bed occupancy processes. This will allow us to derive the steady-state probability of each process in the next section.

5.3.1 Queue Length Process

It is immediate from (5.1) and (5.2) that the process $\{Q_t, D_t | t \in \mathbb{Z}^+\}$ forms a Markov chain. We compute the transition probability of $(Q_t = q_1, D_t = d_1)$ to $(Q_{t+1} = q_2, D_{t+1} = d_2)$, where $(q_i, d_i) \in \{0, 1, \ldots, C_{queue}\} \times \{1, 2, \ldots, 7\}$, $i = 1, 2$. In particular, we have

$$\Pr(Q_{t+1} = q_2, D_{t+1} = d_2 | Q_t = q_1, D_t = d_1) = \Pr(Q_{t+1} = q_2 | Q_t = q_1, D_t = d_1) \quad (5.4)$$

if $d_2 = d_1 + 1 \mod (7)$ and zero otherwise. Now, given Q_t and D_t,

1. $Q_{t+1} = Q_t - S_t + i$, $0 \leq i \leq C_{queue} - Q_t + S_t$, which occurs when the number of surgical arrivals to the queue does not exceed the available capacity of the queue, $C_{queue} - Q_t + S_t$. Such an event occurs with probability $\Pr(A_t^1 = i) = \exp(-\lambda_1)\lambda_1^i / i!$ for any i such that $0 \leq i \leq C_{queue} - Q_t + S_t$.

2. $Q_{t+1} = C_{queue}$, which occurs when the number of surgical arrivals to the queue is greater than $C_{queue} - Q_t + S_t$. This event occurs with probability $\Pr(A_t^1 > j) = 1 - \sum_{i=0}^{j} \exp(-\lambda_1)\lambda_1^i / i!$ where $j = C_{queue} - Q_t - S_t$.

As a result, we have obtained the transition probabilities of (Q_t, D_t).

5.3.2 Bed occupancy Process

According to (5.3), I_{t+1} depends on I_t, S_t, A_t^2, and B_t. We already know that S_t depends on Q_t and D_t, and that A_t^2 is i.i.d. Since the LOS of each patient is assumed to be a geometric random variable with mean $1/\mu$, it possesses the memoryless property. The property implies that each patient on day t, including those who are admitted at the beginning of the day, will continue to stay in the ICU on day $t+1$ with probability $1 - 1/\mu$, regardless of how long he has been staying up until day t. Let $\hat{I}_t$ be the number of busy beds after the admission process of all patients on day t, i.e.,

$$\hat{I}_t = I_t + \min\{S_t, C_{ICU} - I_t\} + \min\left\{A_t^2, C_{ICU} - I_t - \min\{S_t, C_{ICU} - I_t\}\right\}. \quad (5.5)$$

Therefore, conditional on $\hat{I}_t = n$, B_t is a binomial random variable with parameters n and μ such that

$$\Pr(B_t = k|\hat{I}_t = n) = \binom{n}{k}(1-\mu)^{n-k}\mu^k.$$

Notice that B_t depends only on $\hat{I}_t$, which depends on I_t, S_t, and A_t^2. Since $I_{t+1} = \hat{I}_t - B_t$, we conclude that I_{t+1} depends on I_t, Q_t, and D_t. Because (Q_t, D_t) is a Markov chain, it follows that (I_t, Q_t, D_t) is also a Markov chain.

We proceed to discuss the transition probability of $(I_t = i_1, Q_t = q_1, D_t = d_1)$ to $(I_{t+1} = i_2, Q_{t+1} = q_2, D_{t+1} = d_2)$, where $(i_j, q_j, d_j) \in \{0, 1, \ldots, C_{ICU}\} \times \{0, 1, \ldots, C_{queue}\} \times \{1, 2, \ldots, 7\}$, $j = 1, 2$. The transition probability for the case that $d_2 = d_1 + 1 \mod (7)$ can be written as

$$\begin{aligned}
&\Pr(I_{t+1} = i_2, Q_{t+1} = q_2, D_{t+1} = d_2|I_t = i_1, Q_t = q_1, D_t = d_1) \\
&\quad = \Pr(I_{t+1} = i_2|Q_{t+1} = q_2, D_{t+1} = d_2, I_t = i_1, Q_t = q_1, D_t = d_1) \\
&\qquad \cdot \Pr(Q_{t+1} = q_2|Q_t = q_1, D_t = d_1) \\
&\quad = \Pr(I_{t+1} = i_2|I_t = i_1, Q_t = q_1, D_t = d_1)\Pr(Q_{t+1} = q_2|Q_t = q_1, D_t = d_1),
\end{aligned} \quad (5.6)$$

where the last equation follows since I_{t+1} is independent of Q_{t+1}. The transition probability is zero if $d_2 \neq d_1 + 1 \mod (7)$.

Since the analysis of $\Pr(Q_{t+1} = q_2 | Q_t = q_1, D_t = d_1)$ has been done in the previous section, we will focus on investigating the conditional probability $\Pr(I_{t+1} = i_2 | I_t = p_1, Q_t = i_1, D_t = d_1)$. Let us first consider the possible values of $\hat{I}_t$ given I_t, Q_t, and D_t as follows.

1. $\hat{I}_t = C_{ICU}$ if

   - $S_t \geq C_{ICU} - I_t$, implying that the number of surgical arrivals to the ICU is greater than or equal to the number of available beds. Therefore,

   $$\Pr(\hat{I}_t = C_{ICU} | S_t \geq C_{ICU} - I_t, I_t, Q_t, D_t) = 1. \tag{5.7}$$

   - $S_t < C_{ICU} - I_t$ and $A_t^2 \geq C_{ICU} - I_t - S_t$, implying that the number of surgical arrivals to the ICU is less than the number of available beds at the beginning of the day, but that the number of medical arrivals is greater than or equal to the number of available beds after surgical admissions. Therefore,

   $$\begin{aligned} &\Pr(\hat{I}_t = C_{ICU} | S_t < C_{ICU} - I_t, I_t, Q_t, D_t) \\ &\quad = \Pr(A_t^2 \geq C_{ICU} - I_t - S_t | S_t < C_{ICU} - I_t, I_t, Q_t, D_t) \\ &\quad = 1 - \sum_{i=0}^{j-1} \exp(-\lambda_2) \lambda_2^i / i!, \end{aligned} \tag{5.8}$$

   where $j = C_{ICU} - I_t - S_t$.

2. $\hat{I}_t = I_t + S_t + A_t^2 < C_{ICU}$ if $S_t + A_t^2 < C_{ICU} - I_t$, meaning that the sum of surgical and medical arrivals is less than the number of available beds. Therefore, for

$i < C_{ICU}$,

$$\begin{aligned}
&\Pr(\hat{I}_t = i | S_t < C_{ICU} - I_t, I_t, Q_t, D_t) \\
&\quad = \Pr(A_t^2 = i - I_t - S_t | S_t < C_{ICU} - I_t, I_t, Q_t, D_t) \\
&\quad = \exp(-\lambda_2)\lambda_2^j / j!, \qquad (5.9)
\end{aligned}$$

where $j = i - I_t - S_t$.

We have that

$$\begin{aligned}
\Pr(I_{t+1} = i | I_t, Q_t, D_t) &= \Pr(\hat{I}_t - B_t = i | I_t, Q_t, D_t) \\
&= \sum_{j=i}^{C_{ICU}} \Pr(B_t = j - i | \hat{I}_t = j, I_t, Q_t, D_t) \Pr(\hat{I}_t = j | I_t, Q_t, D_t) \\
&= \sum_{j=i}^{C_{ICU}} \Pr(B_t = j - i | \hat{I}_t = j) \Pr(\hat{I}_t = j | I_t, Q_t, D_t).
\end{aligned}$$

The third equation holds since B_t depends only on $\hat{I}_t$. Note that the first term in the sum above is simply the binomial distribution, while the second term has been analyzed in (5.7) and (5.8). As a result,

$$\begin{aligned}
&\Pr(I_{t+1} = i | I_t, Q_t, D_t) \\
&= \begin{cases}
\sum_{j=i}^{C_{ICU}-1} \binom{j}{j-i} (1-\mu)^i \mu^{j-i} \exp(-\lambda_2)\lambda_2^l / l! + & \\
\binom{C_{ICU}}{C_{ICU}-i} (1-\mu)^i \mu^{C_{ICU}-i} (1 - \sum_{k=0}^{m-1} \exp(-\lambda_2)\lambda_2^k / k!), & \text{if } S_t < C_{ICU} - I_t \\
\sum_{j=i}^{C_{ICU}} \binom{j}{j-i} (1-\mu)^i \mu^{j-i}, & \text{if } S_t \geq C_{ICU} - I_t,
\end{cases} \qquad (5.10)
\end{aligned}$$

where $l = j - I_t - S_t$, $m = C_{ICU} - I_t - S_t$. The analysis of the transition probabilities of (I_t, Q_t, D_t) is therefore completed.

5.4 Steady State Probability

5.4.1 Queue Length Process

The stochastic process $\{Q_t, D_t | t \in \mathbb{Z}^+\}$ is a finite-state Markov chain with the size of the state space equal to $7(C_{queue} + 1)$. We denote the probability transition matrix of the chain (Q_t, D_t) by $\mathbf{P}^Q$ such that

$$\mathbf{P}^Q = \begin{pmatrix} \mathbf{0} & \mathbf{Q}^{12} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}^{23} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Q}^{71} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is a $C_{queue} + 1$-dimensional matrix of zeros and $\mathbf{Q}^{ij}$ is the one-step transition matrix of the queue lengths from day of the week i to j, which is the next day of the week adjacent to i. Each element of $\mathbf{Q}^{ij}$ can be expressed as

$$\mathbf{Q}^{ij}_{kl} = \Pr(Q_{t+1} = l, D_{t+1} = j | Q_t = k, D_t = i).$$

We claim that this particular chain has a single recurrent class.

Claim 1. *The Markov chain (Q_t, D_t) has a single recurrent class.*

Proof. We want to show that all states $\{(q, d) : q \in 0, 1, \dots, C_{queue}, d \in 1, 2, \dots, 7\}$ in the Markov chain (Q_t, D_t) are recurrent. Let us fix a state at time t_1 to be (q_1, d_1). Consider another state (q_2, d_2) for which $q_2 \geq q_1$. When $A^1_t = q_2 - q_1 + S_t$, it implies that $Q_{t_1+1} = q_2$. Since such an event occurs with positive probability, $\Pr(Q_{t_1+1} = q_2 | Q_{t_1} = q_1, D_{t_1} = d_1) > 0$. Now, suppose that $Q_{t_1+1} = q_2$. The event $Q_{t_1+2} = q_2$ can occur when $A^1_{t_1+1} = S^1_{t_1+1}$, which also happens with positive probability. It follows that $\Pr(Q_{t_1+2} = q_2 | Q_{t_1+1} = q_2, D_{t_1} = d_1) > 0$. Similarly, $\Pr(Q_{t_1+j+1} = q_2 | Q_{t_1+j} = q_2, D_{t_1} = d_1) > 0$ for any $j > 2$. Therefore, the event that the queue length jumps to q_2 at time $t_1 + 1$ and remains at this value until time $t_2 > t_1 + 1$ where $D_{t_2} = d_2$ has positive probability of occurring. It follows that any

state (q_2, d_2) for which $q_2 \geq q_1$ is accessible from (q_1, d_1).

Now, let us consider the transition from (q_1, d_1) to (q_2, d_2) where $q_2 < q_1$. Since the sum of caps in a week must be strictly positive, there is a positive probability that the weekly number of surgical arrivals is less than the weekly number of surgical departures from the queue. In consequence, the queue length can be decreased at the end of each week with positive probability. Thus, there exist $t_1' > t_1$ and $q_1' \leq q_2$ such that $D_{t_1'} = d_1'$ and $\Pr(Q_{t_1'} = q_1' | Q_{t_1} = q_1, D_{t_1} = d_1) > 0$. This means that the state (q_1', d_1') is accessible from (q_1, d_1). From the previous discussion, we know that (q_2, d_2) is accessible from (q_1', d_1'). This further implies that any state (q_2, d_2) for which $q_2 < q_1$ is also accessible from (q_1, d_1).

Thus, all states are accessible from a fixed (q_1, d_1). By applying the same reasoning to other states than (q_1, d_1), it follows that all states in the Markov chain communicate. The proof is therefore completed. □

For a finite state Markov chain with a single recurrent class, it is known (Gallager [10]) that the stationary distribution vector $\boldsymbol{\pi}$ for which $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^Q$, $\sum_{i=1}^{7(C_{queue}+1)} \pi_i = 1$, and $\boldsymbol{\pi} \geq 0$ is unique. We denote such a vector by $\boldsymbol{\pi}^Q = [\boldsymbol{\pi}^{Q_1}, \ldots, \boldsymbol{\pi}^{Q_7}]$, where the i element of $\boldsymbol{\pi}^{Q_j}$ corresponds to the steady-state probability that there are $i-1$ patients waiting in the queue at the beginning of day j of the week.

5.4.2 Bed Occupancy Process

The process $\{I_t, Q_t, D_t | t \in \mathbb{Z}^+\}$ is a finite-state Markov chain with the size of the state space equal to $7(C_{ICU}+1)(C_{queue}+1)$. Its one-step transition matrix is denoted by $\mathbf{P}^I$:

$$\mathbf{P}^I = \begin{pmatrix} \mathbf{0} & \bar{\mathbf{Q}}^{12} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{Q}}^{23} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{Q}}^{71} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is a $(C_{ICU}+1)(C_{queue}+1)$-dimensional matrix of zeros and

$$\bar{\mathbf{Q}}^{ij} = \begin{pmatrix} \mathbf{I}^{11,ij} & \mathbf{I}^{12,ij} & \dots & \mathbf{I}^{1C_{queue},ij} \\ \mathbf{I}^{21,ij} & \mathbf{I}^{22,ij} & \dots & \mathbf{I}^{2C_{queue},ij} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}^{C_{queue}1,ij} & \mathbf{I}^{C_{queue}2,ij} & \dots & \mathbf{I}^{C_{queue}C_{queue},ij} \end{pmatrix}.$$

$\mathbf{I}^{mn,ij}$ is a matrix of dimension $C_{ICU}+1$, each element of which can be expressed as

$$\mathbf{I}_{kl}^{mn,ij} = \Pr(I_{t+1}=l, Q_{t+1}=n, D_{t+1}=j | I_t=k, Q_t=m, D_t=i).$$

We now show that the Markov chain (I_t, Q_t, D_t) is recurrent.

Claim 2. *The Markov chain (I_t, Q_t, D_t) has a single recurrent class.*

Proof. Fix i_1 for some $i_1 \in \{0,1,\dots,C_{ICU}\}$. On day t, $\hat{I}_t \geq \min\{S_t, C_{ICU}-i_1\}$ implies that B_t can be equal to $\min\{S^1_{t_2-1}, C_{ICU}-i_1\}$ with positive probability. It follows from (5.3) that $I_{t+1} = i_1 + A^2_t = i_2 \geq i_1$ when $A^2_t \leq C_{ICU} - i_1 - \min\{S^1_t, C_{ICU}-i_1\}$, which also occurs with positive probability. Moreover, it is possible that $A^2_t = 0$, which implies that $I_{t+1} = i_1 + \min\{S^2_t, C_{ICU}-i_1\} - B_t = i_2$. It is also possible that B_t can be so large that $i_2 < i_1$. Hence, for any given Q_t and D_t, the transition from $I_t = i_1$ to $I_{t+1} = i_2$ can occur for any $i_2 \in \{1,2,\dots,C_{ICU}\}$. The same argument applies for other values of i_1. It follows that for any given state (I_t, Q_t, D_t), $I_{t+1} = i_2$ for any $i_2 \in \{1,2,\dots,C_{ICU}\}$ with positive probability.

Now, let $I_{t_1} = i_1, Q_{t_1} = q_1, D_{t_1} = d_1$ for some $t_1 \in \mathbb{Z}^+$, $q_1 \in \{0,1,\dots,C_{queue}\}$ and $d_1 \in \{1,2,\dots,7\}$. For any $q_2 \in \{0,1,\dots,C_{queue}\}$ and $d_2 \in \{1,2,\dots,7\}$, there exists $t_2 > t_1$ such that $Q_{t_2} = q_2$ and $D_{t_2} = d_2$ with positive probability because all states in the Markov chain (Q_t, D_t) communicate. From the previous discussion, it follows that, with positive probability, the number of busy beds at time t_1+1 will be i_2 for any $i_2 \in \{1,2,\dots,C_{ICU}\}$ and will remain at this value until time t_2, i.e., $I_{t_2} = i_2$. This implies that any state (i_2, q_2, d_2) is accessible from (i_1, q_1, d_1). By applying the same argument to different values of (i_1, q_1, d_1), it follows that all states in the Markov

chain (I_t, Q_t, D_t) are recurrent, completing the proof. □

Hence, given $\mathbf{P}^I$, there exists a unique stationary probability vector $\boldsymbol{\pi}$ that satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}^I$, $\sum_{i=1}^{7(C_{ICU}+1)(C_{queue}+1)} \pi_i = 1$, and $\boldsymbol{\pi} \geq 0$. We denote such a vector by $\boldsymbol{\pi}^I = [\boldsymbol{\pi}^{I_1}, \boldsymbol{\pi}^{I_2}, \ldots, \boldsymbol{\pi}^{I_7}]$, where $\boldsymbol{\pi}^{I_m} = [\boldsymbol{\pi}^{I_{m,1}}, \boldsymbol{\pi}^{I_{m,2}}, \ldots, \boldsymbol{\pi}^{I_{m,C_{queue}+1}}]$. The i element of $\boldsymbol{\pi}^{I_{m,n}}$ corresponds to the steady-state probability that there are $i-1$ busy beds in the ICU and $n-1$ patients waiting in the queue at the beginning of day m of the week.

5.5 Mean Waiting Time

The average waiting time $\bar{W}$ spent by a customer in the steady state can be computed from Little's law (Gallager [10]), which states that the average number of customers in the system $\bar{L}$ is equal to the arrival rate times $\bar{W}$, i.e., $\bar{L} = \lambda\bar{W}$. In our problem, the system is the queue of surgical patients with the mean number of arrivals per day equal to λ_1. The queueing system is always stable since the queue size is truncated to C_{queue}. By letting Q and D be the queue length and the day of the week in the steady state, we have

$$\begin{aligned}
\bar{L} &= \mathrm{E}[Q] \\
&= \sum_{q=0}^{C_{queue}} q \cdot \Pr(Q = q) \\
&= \sum_{q=0}^{C_{queue}} q \cdot \left(\sum_{d=1}^{7} \Pr(Q = q, D = d) \right) \\
&= \sum_{q=0}^{C_{queue}} \sum_{d=1}^{7} q \cdot \pi_{q+1}^{Q_d}. \qquad (5.11)
\end{aligned}$$

Therefore, $\bar{W}$ can be obtained by applying Little's formula.

5.6 Rejection Rate

Let $\bar{A}$ be the average number of arrivals per day and let $\bar{R}$ be the steady-state average number of rejections per day. We define the rejection rate of the ICU to be the ratio

$\bar{R}/\bar{A}$.

Let $\bar{A} = \bar{A}_1 + \bar{A}_2$ where $\bar{A}_1$ and $\bar{A}_2$ are the average numbers of surgical arrivals to the ICU per day and the average number of medical arrivals to the ICU per day respectively in the steady state. Define S to be the steady-state number of patients sent to the ICU. We have

$$\begin{aligned}
\bar{A}_1 &= \mathrm{E}[S] \\
&= \sum_{i=0}^{C_{queue}} \sum_{j=1}^{7} \min\{i, \mathrm{cap}(j)\} \Pr(Q = i, D = j) \\
&= \sum_{i=0}^{C_{queue}} \sum_{j=1}^{7} \min\{i, \mathrm{cap}(j)\} \pi_{i+1}^{Q_j}. \qquad (5.12)
\end{aligned}$$

Note that the steady-state mean number of medical arrivals per day $\bar{A}_2$ is simply the mean of A_t^2, which is equal to λ_2. Hence, $\bar{A}$ is obtained.

We now proceed to find $\bar{R}$. Let I be the steady-state number of busy beds and let A^2 be the steady-state number of medical arrivals. The rejection of patients can occur in the following events:

1. The number of surgical arrivals to the ICU exceeds the available space in the unit: $S > C_{ICU} - I$. In this event, the number of surgical rejections is equal to $S - (C_{ICU} - I)$. Since medical patients arrive after surgical patients, all medical arrivals are rejected in this case. Let $\bar{R}_1$ be the steady-state average rejections number of this event. We have

$$\bar{R}_1 = \sum_{j=0}^{C_{ICU}} \sum_{i > C_{ICU} - j} ((i - (C_{ICU} - j)) + \lambda_2) \Pr(S = i, I = j). \qquad (5.13)$$

   Note that $\Pr(S = i, I = j) = \Pr\left(\min\{Q, \mathrm{cap}(D)\} = i, I_t = j\right)$ can be computed from the steady-state distribution of the Markov chain (I_t, Q_t, D_t).

2. The number of surgical arrivals to the ICU does not exceed the space in the unit: $S \leq C_{ICU} - I$. In this event, while none of the surgical patients are rejected, $A^2 - C_{ICU} + (I + S)$ medical patients are rejected if $A^2 > C_{ICU} - (I + S)$. Let

$\bar{R}_2$ be the long-run average rejections number of this event, we have that

$$\begin{aligned}\bar{R}_2 &= \sum_{j=0}^{C_{ICU}} \sum_{i \le C_{ICU}-j} \left(\sum_{k > C_{ICU}-(j+i)} (k - C_{ICU} + (j+i)) \cdot \Pr(A^2 = k) \right) \\ &\quad \cdot \Pr(S = i, I = j) \\ &= \sum_{j=0}^{C_{ICU}} \sum_{i \le C_{ICU}-j} \mathrm{E}[A^2 - C_{ICU} + (j+i) | A^2 > C_{ICU} - (j+i)] \\ &\quad \cdot \Pr(S = i, I = j). \end{aligned} \tag{5.14}$$

Each of the expected value can be computed based on the distribution of A_t^2 because A_t^2 is i.i.d. Since the two events are disjoint, $\bar{R} = \bar{R}_1 + \bar{R}_2$. As a result, the long-run average rejections per day is obtained.

5.7 Results and Discussion

We now apply our queueing model to compute performance measures in the ICU with no caps and the uniform cap policy (UCP). We use year 2000 data to parametrize our model. Specifically, $C_{ICU} = 16$, $1/\mu = 3.65$, and the arrival rates are set to be $\lambda_1 = 3.35$ and $\lambda_2 = 2.35$ at a high-utilization regime ($\sim$ 84% utilization) and $\lambda_1 = 2.31$ and $\lambda_2 = 1.42$ at a medium-utilization regime ($\sim$ 72% utilization), according to Section 4.3.3 in Chapter 4. The queue is truncated to $C_{queue} = 30$. The cap is varied from the tightest cap that ensures system stability to no cap. The mean rejection rates and the mean waiting times from the queueing model and simulation at both regimes are shown in Tables 5.1 and 5.2, respectively.

As can be seen, the mean rejection rates computed by the queueing model at different caps in both regimes are close to the results from simulation. The mean waiting times computed by both methods are also consistent with each other. In fact, the mean waiting times from the queueing model are slightly higher since we assume that each surgical patient has to wait at least one day before entering the ICU. Notice, however, that the mean waiting time from the queueing analysis is slightly shorter when the cap is tightest in the medium-utilization regime ([4 4 3 3 3]). This

Cap	Mean rejection rate (%)		Mean waiting time (days)	
	Queueing model	Simulation	Queueing model	Simulation
[5 5 5 5 4]	25.50	25.75	4.19	4.49
[5 5 5 5 5]	25.93	25.82	3.22	2.57
[6 5 5 5 5]	26.02	25.92	2.57	2.01
[6 6 5 5 5]	26.31	26.15	2.22	1.77
[6 6 6 5 5]	26.54	26.27	2.04	1.60
No cap	27.64	26.95	1.43	1.16

Table 5.1: Results from the high-utilization regime

Cap	Mean rejection rate (%)		Mean waiting time (days)	
	Queueing model	Simulation	Queueing model	Simulation
[4 4 3 3 3]	6.88	8.24	5.02	5.01
[4 4 4 3 3]	7.62	8.73	3.28	2.81
[4 4 4 4 3]	8.08	8.93	2.56	2.08
[4 4 4 4 4]	8.27	8.98	2.21	1.66
[5 4 4 4 4]	8.31	9.13	1.94	1.50
No cap	10.03	10.29	1.43	1.04

Table 5.2: Results from the medium-utilization regime

is because our queueing model allows no greater than $C_{queue} = 30$ surgical patients to wait in the queue, while the simulation model is not subject to such limitation. When modeling with simulation, tightest caps tend to build up a relatively much longer queue than this C_{queue}, which potentially leads to the longer mean waiting time of scheduled patients in the simulated system.

In addition, our queueing model exhibits the same trade-off and the impacts of caps on the rejection rate and mean waiting time of the ICU as discussed in Chapter 4. A comparison between the improvement in the rejection rates after applying caps in the medium-utilization regime (3.15%) and in the high-utilization regime (2.14%) also implies that the UC policy performs better in an ICU with moderate workload.

It should be noted that one could extend a discrete-time queueing model to analyze the performance of the service-specific cap policy (SSCP) by introducing the queue length and the bed occupancy processes associated with each type of surgical patients. However, the state-space will grow considerably large to incorporate these additional details, making the numerical computation prohibitive. As a result, we do not make

an effort to analyze the ICU with the SSCP with a discrete-time queueing model.

Chapter 6

State-Dependent Prediction

The development of scheduling policies considered thus far in this thesis has centered around the framework of using *static* caps to control the scheduling of elective surgery cases. In particular, we have demonstrated in Chapter 4 and 5 the trade-off of such caps between the decrease in rejection rates and the increase in mean waiting times of scheduled patients. To achieve a considerable improvement in admission rates, one needs to keep the cap size so small that the backlog of waiting patients could grow considerably large, and so does the mean waiting time. In fact, the results in Chapter 4 have shown that even the tightest cap that still ensures system stability is able to reduce a rejection rate by about 2%, but no more than 3%, at a medium utilization regime ($\sim$ 70%-75%), where the baseline rejection rate before applying caps is about 10%. This still leaves the ICU with a fairly high turnover rate at about 8%. Static cap-based policies thus do not entirely eliminate rejections in the ICU.

In an attempt to further reduce rejection rates, we consider the potential development of *adaptive* cap-based scheduling policies. In this framework, the size of caps is dynamically changed in response to the probabilistic prediction of the future state of the ICU occupancy, which is determined according to current state information. This way, the system receives an early warning signal for times at which the ICU is likely to become overcrowded, and can thereby reduce the cap size accordingly. Similarly, the prediction outcome could indicate the likelihood that system utilization would be low in the future, which allows the ICU administration to raise the cap space

and shorten the waiting times of scheduled surgical patients without sacrificing many service rejections. By properly designing the state-dependent algorithm for adjusting caps, we expect that the policy will be able to improve the admission rate further compared to static caps while maintaining a reasonable amount of waiting time in the ICU.

Motivated by the idea of adaptive caps, we study in this chapter the problem of state-dependent prediction to explore the potential of using current state information in forecasting the future state of the ICU. We now provide the details and scope of the problem to be considered in the chapter.

6.1 Problem Statement

A state of the ICU system consists of several components. All components fall broadly into the following two groups based on the ability of the system to observe them:

- **Perfectly observable components**. This part of the state consists of the components which are observable at the ICU. Examples include the number of current patients in the unit, the length of stay (LOS) up to date of each patient, and the backlog of scheduled patients waiting to enter the ICU.

- **Partially observable components**. This type of state information consists of components that are not directly observed in the system. However, they can be presented in terms of probabilistic outcomes based on perfectly observable components. An important example of these state components is the remaining LOS (or projected departure times) of current patients. Clearly, their exact values are not known to the ICU, but can be inferred probabilistically from the LOS distribution conditional on the LOS up to date, which are perfectly observable. Because of its limited access, partially observable state information presents a challenge in making accurate predictions about the future state. In Section 6.3, we shall see that extra knowledge of this type of state information can have a significant impact on the prediction outcome of the future state of

the ICU.

Given current state information, we are interested in predicting the state of the system at a future point in time. The prediction outcome is measured in the form of probability distributions, and can be achieved by simulating many sample paths that initially start with the same given current state. Of all future state components, the one of particular interest is the number of busy beds, since it directly implies whether the ICU is likely to reject patients or will be under-utilized. As a result, our goal is to determine the distribution of the future number of busy beds based on current state information.

The study of state-dependent predictions in this thesis is divided into two main parts based on the access to partially observable state components, in particular the remaining LOS of each patient. In the first part, the problem is built on the assumption that the remaining LOS of each patient is unknown to the ICU and that the only information that can be used to estimate it is the current LOS. In the second part, we assume that the ICU staff has additional knowledge that allows them to know in advance whether or not each current patient will leave the unit before (or after) a given time in the future. We incorporate this extra information into the study of state-dependent predictions in the second part.

6.2 Part I: State-Dependent Prediction Based on Perfectly Observable State Information

In steady state, the estimate of the probability that an ICU will be in a particular state at a particular time can be obtained by means of simulation. For example, the stationary probability that the ICU is fully-occupied at a given time t is estimated by simulating many sample paths of the system in steady state, then calculating the frequency with which all beds are filled at t.

However, when conditional on a current state, it is likely that the system, which could have already been in steady state, will start afresh. As the system renews and

now starts with the given current state, we expect that the probability distribution of states at the near future, instead of following the stationary distribution, would highly depend on this current state information. As the time difference between the present and future becomes larger, we expect that the impact of the current state information on the future state will gradually disappear. In other words, the likelihood of the state of the system at a distant future time would start to be less dependent on the current state and tend to occur with the probability approximately equal to that of a steady-state system.

As such, we believe that the knowledge of a current state would be informative in predicting the behavior of the ICU at the near future. Our goal in this section is to

1) investigate the impact of current state information on the future state of the ICU, and

2) to study the fading in the impact of current states on future states as the time difference between the present and future grows.

We now describe a solution method to achieve these goals.

6.2.1 Notation

T_c Current time.

T_p Time upon which a prediction is made. $T_p > T_c$.

τ $\tau = T_p - T_c$.

N_c Number of busy beds at T_c.

N_p Number of busy beds at T_p.

B_c Number of surgical patients who are currently waiting for surgery.

L Vector of actual LOS.

L_c Vector of LOS up to T_c.

L_r Vector of the remaining LOS of patients in the ICU at T_c.

6.2.2 Methods

Let I_c be the set of all observable state information at time T_c. In particular, I_c consists of T_c, N_c, B_c, and L_c, all of which are deterministically known at time T_c. More precisely, all state components except T_c are random variables, and so is the set I_c. These random variables become deterministic values once they are observed at T_c, and we denote generically these deterministic state components by i_c. Given L_c, L_r for each patient is a random variable and is generated from the empirical distribution of the LOS for a given service conditional on L_c.

To see the impact of I_c on the unit occupancy at a future point in time T_p, we are interested in computing the conditional probability that the ICU will become full at T_p, which is denoted by $P_c = \Pr(N_p = C_{ICU} | I_c = i_c)$, for various sets of current state i_c. This involves the following two steps:

- **Step I: Generate a current state** i_c. We generate a set i_c at T_c in a steady-state system by simulating an ICU starting at a long-distant past up to T_c and recording the state information at this time. This i_c will be used as a current state in estimating the corresponding P_c.

- **Step II: Estimate** P_c **for a predefined** i_c. Given i_c from the previous step, many samples paths, each of which starts at T_c with i_c, are simulated up until T_p. L_r in each sample path is randomly generated from the respective LOS distribution conditional on L_c. Then, the value P_c associated with this i_c is estimated from the frequency with which the simulated ICU is full at time T_p.

As a result, we have obtained many records of P_c at a given T_p based on different i_c. The same approaches can be used to estimate a set of P_c at different T_p by simply varying its values.

To present the results from simulation, we plot the distribution of P_c for each fixed $\tau = T_p - T_c$. Since P_c is a random variable of I_c and each deterministic i_c is high-dimensional in that the set contains more than one component of state information, it is hard to present the distribution of P_c with respect to many states i_c. To avoid this difficulty, we choose to plot the probability distribution of P_c as a function of its

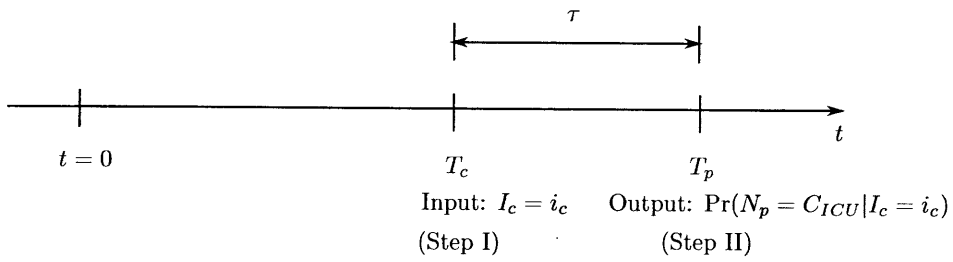


Figure 6-1: Time diagram of the state-dependent prediction problem in Section 6.2

own values. That is, the distribution will be given as a frequency with which each value of P_c occurs in the simulation.

6.2.3 Results and Discussion

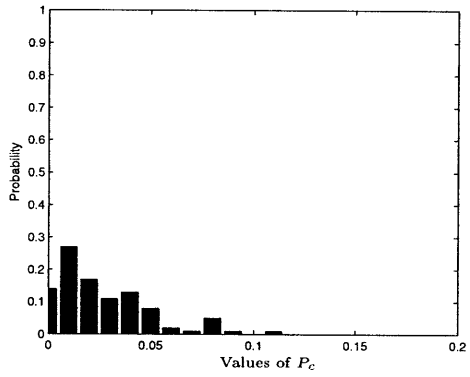
When τ is small, we expect to see a wide range in the distribution of P_c. This follows from our speculation that different current states i_c tend to result in different probabilities that the ICU will be full in the near future. As τ grows, we believe that the distribution of P_c would become less variable, meaning that the system starts to reach stationarity and its state at the distant future would be less correlated to the state at the present time.

Now, we provide computation results and discussion based on the state-dependent prediction problem of this part. The ICU data and environment in 2008 is used in every computational experiment of this chapter. Specifically, the number of beds is fixed throughout at 28. The arrival rates of surgical and medical patients, which have been uncensored according to the method described in Section 3.2.2 of Chapter 3, are provided in Table 6.6 and Table 6.7 respectively at the end of this chapter. Note that the rates of surgical arrivals provided in Table 6.6 are computed to be proportional to the number of weekdays in 2000, and they are higher than the rates in Table 2.5 of Chapter 2, which are weighted by the total number of days in 2008. Also, the arrival rates of medical patients only include those who arrived to the main 28-bed unit since we have no access to the data of the 10-bed medical ICU. The LOS data from 2003

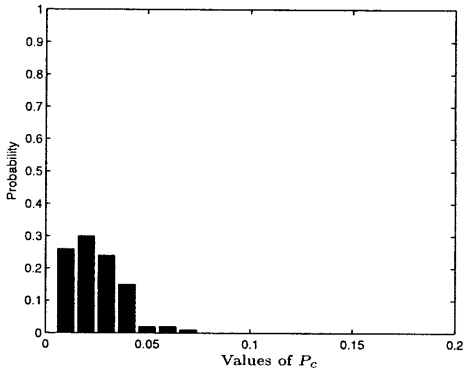
to 2007 is used to generate the LOS of patients in the simulation. Elective surgery patients are scheduled to the ICU based on the uniform cap policy (UCP) with the fixed cap [7 7 6 6 6].

To determine the capacity of the secondary ICU (SICU), we calibrate our simulation model to give an estimate for total rejection rate of surgical patients that is close to the actual diversion rate of surgical patients in 2008, which is at 3.56%. When the number of beds in the SICU is set to four, which is used in the model for the ICU in 2000 (see Section 3.3.2 in Chapter 3), the rejection rate of surgical patients from simulation is equal to 4.91%. We then reduce the capacity of the SICU to close the gap between the surgical rejection rate from simulation and the actual one. It turns out that we are able to obtain the closest estimate at 3.76% when no single beds are allocated to the SICU in simulation. This implies that the ICU capacity in 2008 was able to match the demand for critical care on its own, so the unit became rarely overcrowded, and diverting surgical patients to create space for medical patients was an uncommon activity as a result. We thus set the capacity of the SICU to zero in our simulation model for the ICU in 2008.

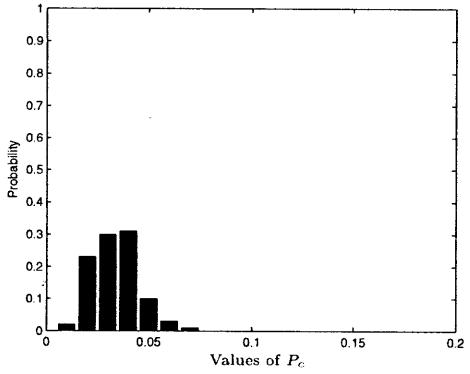
Following this set-up, we randomly generate I_c for 100 times for each fixed τ, which gives us 100 values of P_c. T_c is set to be the beginning of a Friday. For each I_c, the corresponding P_c is estimated by simulating 1000 sample paths. Fig.6-2 illustrates the probability distributions of P_c with respect to its values at different τ. The CV of P_c at different τ is shown in Fig.6-3. The steady-state unconditional probability that the ICU is full at the beginning of a Friday is computed to be 0.044.



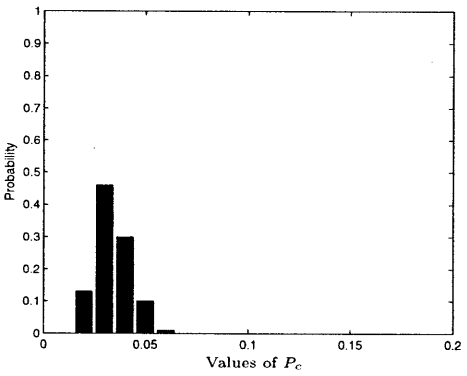
(a) $\tau = 1$ week



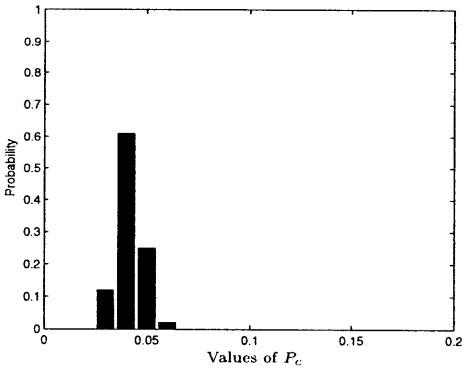
(b) $\tau = 2$ weeks



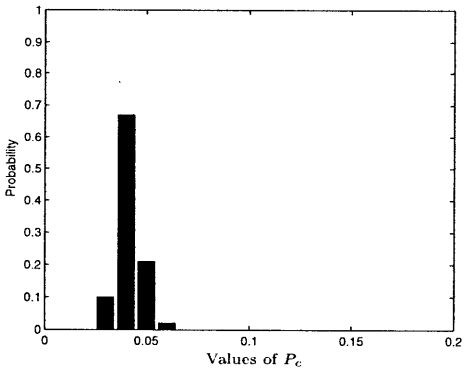
(c) $\tau = 3$ weeks



(d) $\tau = 4$ weeks

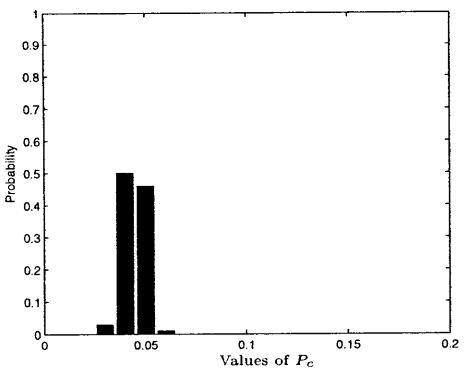


(e) $\tau = 8$ weeks

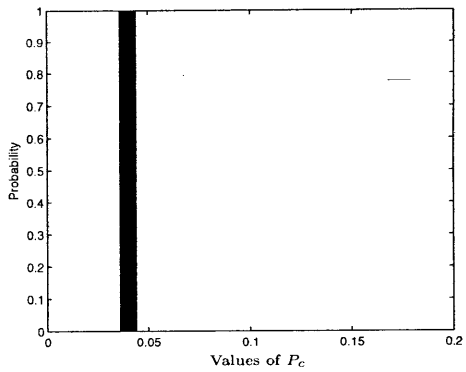


(f) $\tau = 16$ weeks

Figure 6-2



(g) τ = 24 week



(h) τ = 48 weeks

Figure 6-2: Distributions of P_c with respect to its values at different τ

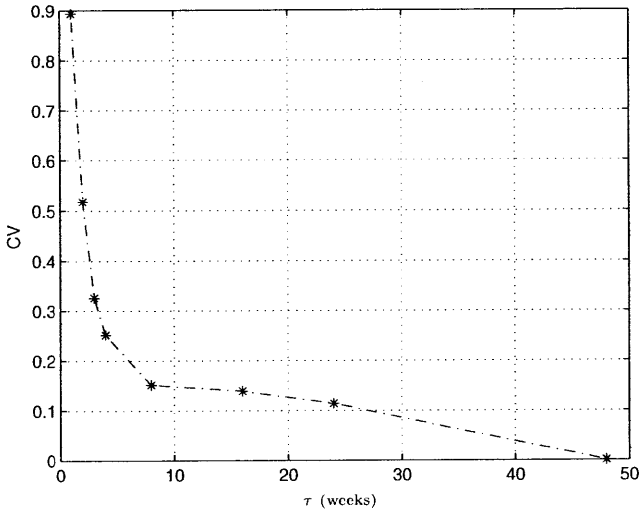


Figure 6-3: CV of P_c at different τ

The results shown in both figures are consistent with our conjecture. As can be seen in Fig.6-2, the distributions of P_c in the first few weeks after the current time T_c are relatively more variable, which indicates that the state of the system in the near future are likely to be dependent on a current state I_c. As τ grows, the distributions of P_c now become less variable and gradually converge to the unconditional stationary probability (0.044). Fig.6-3 also shows that P_c tends to become less variable along with the increase in τ, as evidenced by the decrease in its CV. These results clearly

demonstrate the diminishing effect of current state information on the state of the ICU at the distant future.

6.3 Part II: State-Dependent Prediction Assuming Additional Information on the Departure Times of Current Patients

In the previous section, our state-dependent prediction problem is conservative about the information on the remaining LOS (or the projected departure times) of current patients in the ICU, as it does not assume any knowledge to estimate this component of state information. In reality, however, doctors are often able to make an educated guess, based on their experience, about the projected departure times of current ICU patients after monitoring their health conditions for a certain period of time. For instance, the ICU can identify if a patient who was admitted two days ago would leave the unit within one week afterwards. Currently, the ICU at CHB does not attempt to estimate the departure times of its current patients. Our goal is to determine if it is worthwhile for the unit to make this effort. That is, we shall investigate whether knowing the remaining LOS in advance can bring value to the prediction of the future state of the ICU.

Toward this end, we study in this part state-dependent predictions that incorporate the ability to predict the departure times of existing patients. We now formulate the problem to be considered in this section.

6.3.1 Problem Formulation

The same notation as listed in Part I will be used in this section. Suppose that the system is now at time T_c with the corresponding perfectly-observable current state being $I_c = i_c$, the goal of the state-dependent prediction problem is to estimate the conditional probability that the system will be full at time T_p. Let us denote this quantity by $\tilde{P}_c$ for notational convenience.

The only assumption we made about the remaining LOS L_r is that, for every patient currently in the ICU, it can be predicted whether he is going to leave the unit before or after a fixed time T_d. The exact departure times, however, are not assumed to be explicitly known in our study. This assumption on L_r divides the current pool of N_c ICU patients into two groups of N_c^s and N_c^l patients, $N_c^s + N_c^l = N_c$, such that

- N_c^s of them will leave during $(T_c, T_d]$. Equivalently, their $L_r \in (0, T_d]$.
- N_c^l of them will leave after T_d. Equivalently, their $L_r \in (T_d, \infty)$.

In our problem, each patient is assigned to leave before or after T_d based on the probability that his L_r is greater than $T_d - T_c$ conditional on his current LOS L_c, $\Pr(L \geq L_c + (T_d - T_c) | L > L_c)$. In particular, the first N_c^l patients whose LOS L are most likely to go beyond the next $T_d - T_c$ days are assumed to depart after T_d. The rest N_c^s patients are conditioned to leave before T_p. The LOS L of these N_c^l patients are generated by the empirical LOS distribution conditional on $L \in (L_c, L_c + (T_d - T_c)]$ according to their seasonality and services.

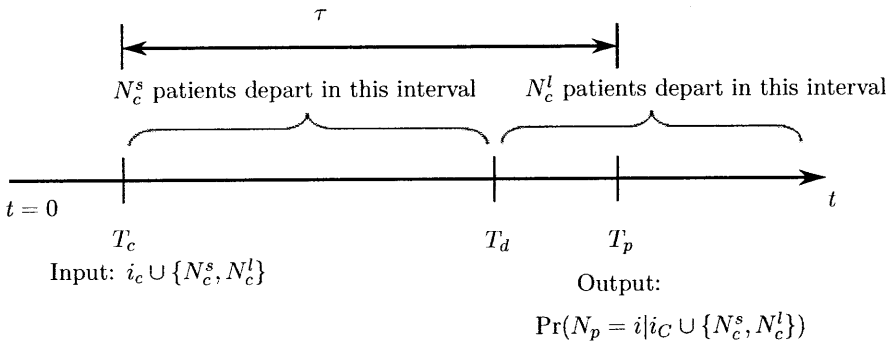


Figure 6-4: Time diagram of the state-dependent prediction problem in Section 6.3

By integrating the extra information on L_r, the original problem is recast to find the probability $\tilde{P}_c$ that the ICU is full at a fixed T_p conditional on $I_c = i_c$, N_c^s, and N_c^l. To do so, we generate many sample paths of patients arriving to the ICU starting at T_c up to T_p with the initial state being set according to the given i_c, N_c^s, and N_c^l. L_r of each patient is generated as discussed earlier. We then simulate these sample

paths by the simulation model and record the frequency with which $N_p = C_{ICU}$, which gives us the estimate of $\tilde{P}_c$.

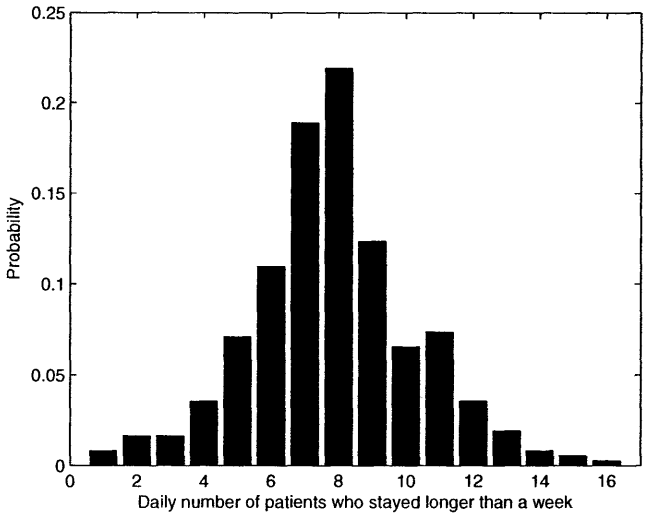


Figure 6-5: Distribution of the daily number of patients who stayed longer than a week in 2008.

6.3.2 Results and Discussion

As the extra information about L_r can be translated into the state components N_c^s and N_c^l, we are now interested in understanding the impact of these two state parameters on the future probability that the ICU becomes full, $\tilde{P}_c$. Our expectation is that the $\tilde{P}_c$ could increase as N_c^l increases (more long-stay patients in the current system). In addition, as the prediction time horizon τ grows while T_d is fixed, we expect that the impact of N_c^l on the future system overcrowding probability might disappear similar to the results from Part I.

The same ICU data and environment as in Section 6.2.3 is used to generate and simulate sample paths in this section. To see the impacts of N_c^s and N_c^l on the prediction outcomes, we vary these two parameters once at a time while fixing all other components of state information. In particular, we fix T_c to be on Friday at 12:00 AM and $B_c = 15$. A set of the LOS up to date L_c is randomly generated by

simulating a sample path and recording the corresponding LOS at T_c. It is important to note that T_d is fixed to be one week throughout this section. The distribution of the daily number of patients who stayed longer than a week in 2008 is provided in Fig.6-5. The true values of N_c^l (ranging from 1-16) as shown in the figure will be used in each of our experiments. Finally, the conditional probability $\tilde{P}_c$ is estimated from simulating 5000 independent sample paths starting with given i_c, N_c^s, and N_c^l from T_c to T_p. We now provide computational results and discussion based on the formulation of the state-dependent prediction problem of this part.

When $T_p = T_d$: τ = 1 week and $T_d = T_c + 7$ days

The prediction time horizon τ is fixed at one week, which means we set T_p to be the same time as T_d. We first investigate the impact of the number of long-stay patients N_c^l on $\tilde{P}_c$ by varying this current state information while fixing the total number of current patients N_c. Table 6.1 and Fig.6-6 summarize the values of $\tilde{P}_c$ associated with the varying N_c^l at the fixed $N_c = 20$. As can be seen, a larger number of current long-stay patients in the ICU leads to the increase in $\tilde{P}_c$. In particular, Fig.6-6 implies that the future overcrowding probability at T_p grows non-linearly with respect to N_c^l.

Parameter	$N_c^l = 2$	$N_c^l = 4$	$N_c^l = 6$	$N_c^l = 8$	$N_c^l = 10$	$N_c^l = 12$	$N_c^l = 14$
$\tilde{P}_c$	0.0000	0.0006	0.0050	0.0172	0.0488	0.1094	0.2044

Table 6.1: $\tilde{P}_c$ at different values of N_c^l when $N_c = 20$ and τ = 1 week

N_c^l	$\tilde{P}_c$						
	$N_c = 14$	$N_c = 16$	$N_c = 18$	$N_c = 20$	$N_c = 22$	$N_c = 24$	$N_c = 26$
8	0.0192	0.0192	0.0186	0.0172	0.0162	0.0150	0.0136
12	0.1122	0.1104	0.1106	0.1094	0.1050	0.0958	0.0824

Table 6.2: $\tilde{P}_c$ at different values of N_c when $N_c^l = 8$ and 12 and τ = 1 week

Fixing N_c^l, we now study the impacts of varying N_c (or N_c^s) on the values of $\tilde{P}_c$. Table 6.2 and Fig.6-7 present the results of $\tilde{P}_c$ based on different values of N_c while fixing N_c^l to be 8 and 12, respectively. Observe that the increase in N_c leads to the

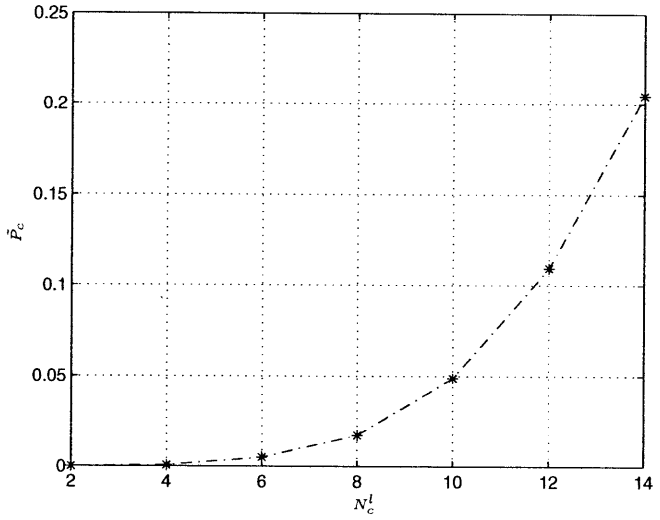


Figure 6-6: $\tilde{P}_c$ at different values of N_c^l when $N_c = 20$ and $\tau = 1$ week

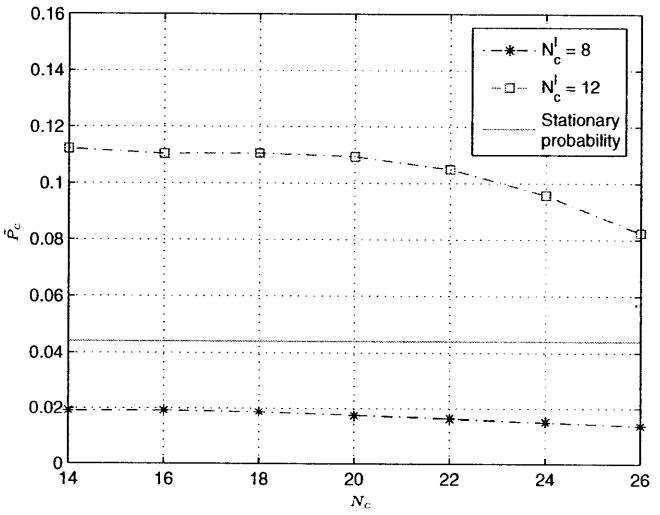


Figure 6-7: $\tilde{P}_c$ at different values of N_c when $N_c^l = 8$ and 12 and $\tau = 1$ week

decrease in $\tilde{P}_c$. This is because the ICU with more current busy beds (larger N_c) has less space and tends to reject more newcoming patients during the interval $(T_c, T_p]$. It is then less likely that the new arrivals who are going to stay beyond T_p would be admitted to the ICU during this interval. As a result, the probability that the ICU will be full at T_p becomes lower in this case. On the other hand, a smaller N_c provides

more space for the ICU to admit patients during $(T_c, T_p]$. This increases the chance for long-stay patients to enter the unit during this period, which in turn leads to a higher $\tilde{P}_c$. Note that we expect the service-specific cap policy (SSCP) to potentially play a role in controlling the admission of patients with long LOS into the ICU.

Despite the increase in $\tilde{P}_c$ from reducing N_c at a fixed N_c^l, we want to emphasize that this change is insignificant compared to the change in $\tilde{P}_c$ when the value of N_c^l is varied. This result allows us to conclude that the estimate for the probability that the ICU will be fully-occupied at a future point in time is practically independent of the total number of current patients, but depends almost entirely on the number of current patients who will stay in the unit pass that time.

When $T_P > T_d$: $\tau > 1$ week while $T_d = T_c + 7$ days

Elective surgical patients in the actual ICU are always scheduled in advance by block-based scheduling processes. In this case, the recognition for the unit occupancy at only a week away might fail to provide enough time for the ICU to adjust their current admission policy in order to prevent undesirable events (such as system overcrowding or under-utilization) that would likely occur next week. For this reason, we are interested in studying the impact of current state information on the state of the ICU in the farther future.

As a result, the prediction time span τ in the following experiments is extended to be longer than a week. However, we limit the time frame within which the departures of each current patient can be identified to be one week, meaning that T_d is still fixed at $T_c + 7$ and $T_d < T_p$. This is because the ICU staff might not be able to make accurate guesses on whether each of their current patients would depart after the very distant future.

Fixing N_c, we now investigate the impacts of N_c^l on $\tilde{P}_c$ at various prediction time horizons τ. Table 6.3 and Fig.6-8 show the results of $\tilde{P}_c$ that correspond to different N_c^l at $\tau = 1$, 2, and 4 weeks, while fixing $N_c = 20$. As can be seen, $\tilde{P}_c$ starts to be less sensitive to the change in N_c^l as τ grows even though it still becomes larger along with the increase in N_c^l. This is because the ability to identify the departure times

of current patients is restricted to be within the period of one week, which in some way causes the state of the system after one week in the future to become gradually less dependent on the current knowledge of N_c^l and N_c^s. Notice that when $\tau = 2$ weeks $\tilde{P}_c$ still increases non-linearly with respect to N_c^l, but with a slower growing rate compared to when $\tau = 1$ week. When τ is extended to 4 weeks, we clearly observe the diminishing impact of current state information on the future state, as all the values of $\tilde{P}_c$ at different N_c^l shown in Table 6.3 become relatively close to the steady-state probability 0.044. In fact, the relationship between $\tilde{P}_c$ and N_c^l at $\tau = 4$ weeks in Fig.6-8 tends to be linear. This finding is consistent with the results observed in Part I, which suggest that current state information could be useful in forecasting a future state up to the first few weeks after the current time.

τ	$\tilde{P}_c$						
	$N_c^l = 2$	$N_c^l = 4$	$N_c^l = 6$	$N_c^l = 8$	$N_c^l = 10$	$N_c^l = 12$	$N_c^l = 14$
1 week	0.0000	0.0006	0.005	0.0172	0.0488	0.1094	0.2044
2 weeks	0.0046	0.0088	0.0190	0.0374	0.0532	0.0782	0.1152
4 weeks	0.0208	0.0268	0.0344	0.0366	0.0478	0.0546	0.0612

Table 6.3: $\tilde{P}_c$ at different values of N_c^l when $N_c = 20$ and $\tau = 1$, 2, and 4 weeks

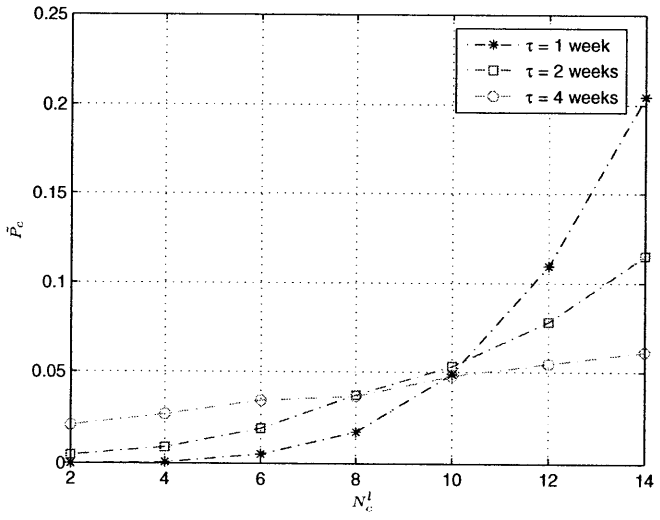


Figure 6-8: $\tilde{P}_c$ at different values of N_c^l when $N_c = 20$ and $\tau = 1$, 2, and 4 weeks

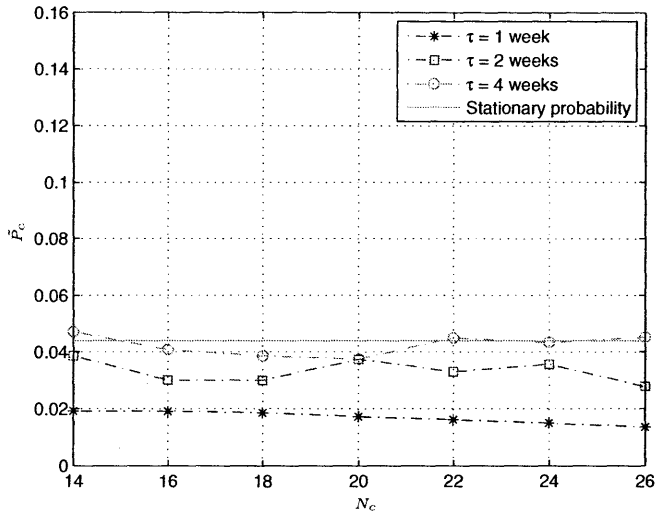
τ	N_c^l	$\tilde{P}_c$						
		$N_c = 14$	$N_c = 16$	$N_c = 18$	$N_c = 20$	$N_c = 22$	$N_c = 24$	$N_c = 26$
1 week	8	0.0192	0.0192	0.0186	0.0172	0.0162	0.0150	0.0136
	12	0.1122	0.1104	0.1106	0.1094	0.1050	0.0958	0.0824
2 weeks	8	0.0386	0.0300	0.0300	0.0374	0.0330	0.0356	0.0278
	12	0.0712	0.0834	0.0736	0.0782	0.0720	0.0782	0.0672
4 weeks	8	0.0472	0.0408	0.0386	0.0374	0.0450	0.0434	0.0452
	12	0.0546	0.0518	0.0554	0.0546	0.0534	0.0586	0.0574

Table 6.4: $\tilde{P}_c$ at different values of N_c when $N_c^l = 8$ and 12 and $\tau = 1$, 2, and 4 weeks

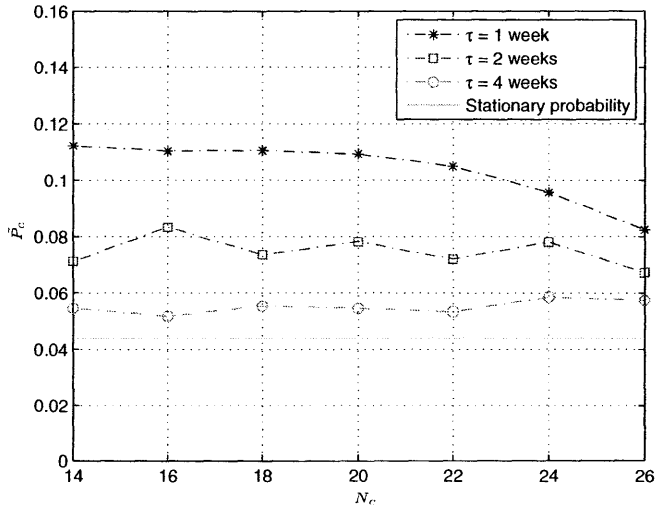
Let us now fix the number of current patients N_c^l who are staying in the ICU through the next week to study the impacts of the the total number of current patients N_c on the system overcrowding probability $\tilde{P}_c$ at $T_p > T_d$. Table 6.4 and Fig.6-9 show the results of $\tilde{P}_c$ with respect to the varying N_c at $\tau = 1$, 2, and 4 weeks, while N_c^l is fixed to be 8 and 12, respectively. As can be seen in Fig.6-9, when $\tau > 1$ ($T_p > T_d$), increasing N_c does not always lead to the lower values of $\tilde{P}_c$ as in the case where $\tau = 1$ week ($T_d = T_p$). This implies that the system overcrowding probability at T_p can be dependent on other current state information as well. In fact, although we already know in advance that N_c^l patients will definitely depart after T_d, we do not know how many of these N_c^l patients would depart after T_p when $T_p > T_d$. It turns out that the probability that each of the N_c^l current patients will stay pass T_p can have a non-trivial impact on $\tilde{P}_c$, the likelihood that the ICU is full at T_p.

For this reason, it appears that states which differ in components other than N_c^l can still lead to the significantly different values of $\tilde{P}_c$ in the case where $T_p > T_d$ and N_c^l is fixed. Let us provide a concrete example to support this claim. In particular, we set up two extreme simulation scenarios, with $N_c^l = 10$, $\tau = 2$ weeks, and all other state information being equal except for the members in the group of N_c^l patients, who are assumed to leave after a week. The first scenario assumes that all N_c^l patients are from Group 1[1] and Group 2[2] of surgical patients, which have short and intermediate LOS on average, respectively. The second scenario assumes all N_c^l patients to be

[1]Short LOS patients: neurosurgical, ORL, plastics surgery, urology, and OMFS patients
[2]Intermediate LOS patients: orthopedic, trauma, and IntRadio patients



(a) $N_c^l = 8$



(b) $N_c^l = 12$

Figure 6-9: $\tilde{P}_c$ at different values of N_c when $N_c^l = 8$ and 12 and $\tau = 1$, 2, and 4 weeks

general surgery and medical patients, which tend to stay in the ICU for a long period of time on average. Clearly, the N_c^l patients in the second scenario are more likely to stay longer than two weeks compared to those N_c^l patients in the first scenario. The value of $\tilde{P}_c$ computed from the first scenario is 0.0144, while the second scenario gives

the estimate of 0.0518 for $\tilde{P}_c$. As can be seen, there is a huge difference in the chances of system overcrowding at $T_p > T_d$ as a result of different current states, despite the same N_c^l.

Thus, we conclude that, when the time of prediction T_p is farther than the time T_d at which the ICU could anticipate the departures of their current patients, the probability that the unit will be full at T_p can possibly depend on components of current state information other than N_c^l and N_c, which are the current number of patients who are known beforehand to leave the ICU after T_d and the total number of patients at T_c, respectively.

6.3.3 Unit Occupancy as a Function of the Number of Long-Stay Patients

In this section, we further investigate the relationship between the number of patients who stayed longer than a week in each day and the unit occupancy in the future based on actual data rather than simulation experiments. Our goal is to see whether knowing the number of long-stay patients would be helpful in determining the future state of the ICU in reality. To gather the data for this analysis, we proceed as follows:

1. Consider a day in the calendar year. At the beginning of this day, we count the number of patients who stayed in the ICU longer than a week after this day from the data. For brevity, let us regard this quantity as the number of current long-stay patients.

2. Consider a fixed τ. With respect to this number of current long-stay patients, we record the corresponding number of busy beds (or the unit occupancy) at τ weeks after this day from the data.

These two steps give us one data point, which consists of the number of current long-stay patients and the corresponding unit occupancy at τ weeks ahead. We then use this to collect 730 data points in total based on the ICU census from 2007 -2008. We

choose to analyze the data from these two years because it was in 2007 that the unit capacity was raised to 28 beds (plus one crash bed).

Parameter	$\tau = 1$ week	$\tau = 2$ weeks	$\tau = 4$ weeks
R^2	0.2942	0.1486	0.0366

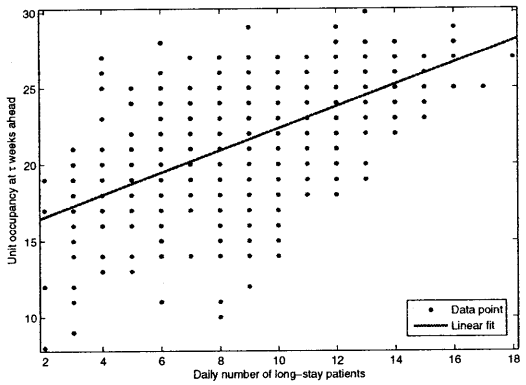
Table 6.5: Values of R^2 of the linear regressions in Fig.6-10

The method of linear regression is used to model the relationship between the future unit occupancy, which is a dependent variable, and the number of current long-stay patients, which is the only independent variable (or a regressor). The model takes form $f = a_1 x + a_2$, where in our context f is the unit occupancy at τ weeks ahead and x is the number of current long-stay patients. Fig.6-10 shows sets of data points and the corresponding linear fits when $\tau = 1, 2$, and 4 weeks, respectively. Note that the unit occupancy can exceed the maximum ICU capacity at 28 since the data includes the records of those surgical patients who were cared outside the ICU during the considered period. Table 6.5 gives the coefficient of determination, R^2, of the linear regressions in Fig.6-10. The coefficient R^2 can be used as an indicator for the goodness of fit of a regression model. For a given data set (x_i, y_i) and an associated modeled value f_i from a regression, R^2 is defined as

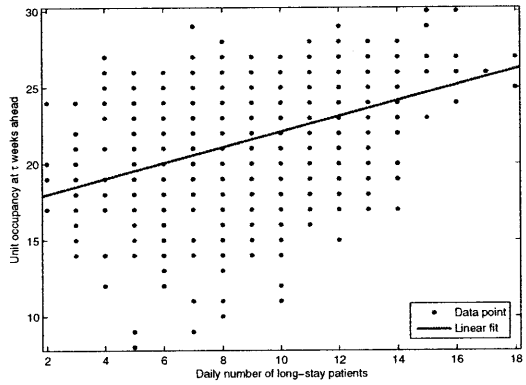
$$R^2 \equiv 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where $0 \leq R^2 \leq 1$ and $\bar{y}$ is the mean of y_i. The high value of R^2 statistically implies that a large portion of the dependent variable (future unit occupancy) is explained by the independent variable (the number of long-stay patients). Clearly, when a regression line perfectly fits a given set of data points, the numerator $\sum_i (y_i - f_i)^2$ becomes zero and R^2 is equal to one.

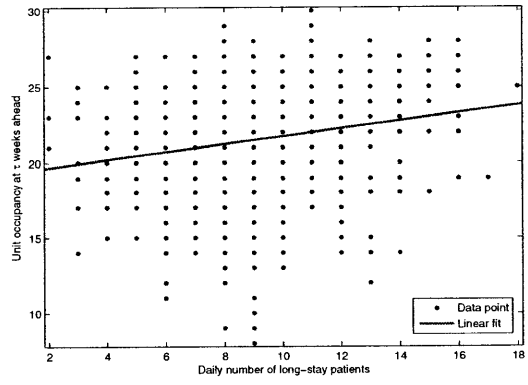
As can be seen in Fig.6-10, the information on the current number of long-stay patients can have a significant impact on the "real" unit occupancy one week away from the present. In fact, when the number of current long-stay patients is greater than 14, the regression analysis at $\tau = 1$ week indicates that the ICU tends to become



(a) $\tau = 1$ week: $a_1 = 0.7204$, $a_2 = 15.14$



(b) $\tau = 2$ weeks: $a_1 = 0.5134$, $a_2 = 16.96$



(c) $\tau = 4$ weeks: $a_1 = 0.2569$, $a_2 = 19.17$

Figure 6-10: Data of the unit occupancy at $\tau = 1, 2$, and 4 weeks ahead with respect to the number of current patients who stayed longer than one week in 2007-2008 and the linear regression of the respective data sets

very busy, as the number of busy beds in the week ahead exceeds 25. The slope of the linear fit then becomes smaller as the value of τ grows larger than one week, which implies that knowing the precise number of current patients who would stay longer than a week becomes less useful to the prediction of the unit occupancy in the relatively farther future. These results are consistent with the results from the simulation-based study of state-dependent predictions in Section 6.3. In addition, Table 6.5 shows that the goodness of fits, as indicated by the coefficient R^2, declines as τ grows. This implies that the unit occupancy in the farther future becomes harder to predict from knowing only the number of current patients who will stay longer than a week.

6.3.4 Concluding Remarks

The results in this section demonstrate the importance of knowing the patients' remaining LOS on the prediction of the future state of the system. This information allows the ICU to compute the chance of system overcrowding at a future point in time by simply looking at the number of current patients who are going to depart after that time. As such, we believe that it is advisable for the ICU staff to make an effort to estimate the remaining LOS of their current patients. The question remains is how accurate and within what length of the time window the doctors are able to predict the departures of patients. This can be an interesting subject of study for the hospital, and its results can be useful for constructing computational scenarios that integrate more realistic observation of the remaining LOS into the state-dependent prediction problem.

More importantly, the ability to forecast the state of the ICU based on current state information suggests the potential of developing adaptive cap-based admission control polices that consistently adjust caps based on the feedback from state-dependent predictions to counteract against any future undesirable effects, such as system congestion or under-utilization. A typical algorithm would be to lower caps whenever the prediction outcome indicates a likely sign of system overcrowding to lower the likelihood of future rejections, and to raise caps when the prediction is the

opposite in order to prevent under-utilization of ICU resources as well as to reduce waiting times of scheduled patients. Although the development of such policies is beyond the scope of this thesis, we believe that an adaptive cap-based policy, when properly designed, should be capable of overcoming the fundamental tradeoff between rejection rates and mean waiting times presented by static cap-based policies. That is, the policy would be able to reduce the rejection rate in the ICU further than any particular set of static caps, while still keeping the mean waiting time under control and thus ensuring the stability of the ICU system.

Arrival Rates in 2008

Type of surgical patients	Winter arrival rate	Non-winter arrival rate
Neurosurgical	1.08	1.22
ORL	1.13	1.53
Plastics	0.15	0.38
Urology	0.05	0.05
OMFS	0.13	0.16
Orthopedic	0.56	0.81
Trauma	0.08	0.11
IntRadio	0.14	0.16
General surgery	1.21	1.35
Other surgery	0.05	0.10
Sum	4.57	5.89

Table 6.6: Daily arrival rates of surgical patients in 2008

Season	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Winter	2.17	2.47	1.95	2.06	2.36	2.23	3.61
Non-winter	1.52	1.29	1.39	1.32	2.21	1.28	1.76

Table 6.7: Daily arrival rates of medical patients in 2008

Chapter 7

Conclusions

In this thesis, we developed a simulation framework for the ICU at CHB and used it to study the impacts of various admission control policies on the ICU. We also used simulation methodology for the purposes of statistical forecasting of the state of the ICU system. As a first step, we performed an extensive statistical analysis of the ICU data, which led to significant insight into the arrivals and length of stay (LOS) of patients statistics. We then built a simulation model for the ICU, based on the results from the statistical analysis as well as the policies and practices in the real ICU. The model was validated to provide accurate estimates for several performance metrics such as rejection rates in the ICU.

The admission control policies considered involve the use of daily caps to control the number of elective surgery cases that can be scheduled on a single day. The first one is the uniform cap policy (UCP), which is the existing policy in the ICU at CHB. This policy enforces caps on the number of total elective patients allowed per day in order to reduce variability in demand for an ICU from these patients. By utilizing the service-based heterogeneity in the LOS of surgical patients, we also considered the service-specific cap policy (SSCP), which uses separate caps to control the admission of separate groups of elective surgery patients based on their average LOS.

We investigated the performance of these two cap-based policies by using the simulation model. The UCP was shown to be capable of smoothing the demand from elective surgeries and reducing the rejection rates in the ICU. These improvements

are achieved at the expense of the increase in the mean waiting time of scheduled patients. Compared to the UCP, the SSCP further reduces variability in scheduled surgical demand and lowers the rejection rate, but it also further increases the mean waiting time. Both cap-based policies were shown to be most effective when the system utilization is around $70\% - 75\%$. At best, the UCP and SSCP decrease the rejection rate up to 2% and 2.2%, respectively. We also showed that the rejection rate in the ICU can be further reduced by discharging its patients earlier (or reducing their LOS), and this benefit is evident even when patients are assumed to depart just a few hours earlier on average. This observation suggests more frequent monitoring ICU patients for their potential discharge as well as an attempt to remove ICU outflow obstructions.

A discrete-time queueing model was developed to analyze the patient flow in the ICU at CHB. The model was shown to provide estimates for the rejection rate and mean waiting time in the ICU with the UCP that are consistent with the simulation results.

We introduced the notion of state-dependent prediction, which aims to identify the probability with which a particular state of the ICU would occur in the future based on the current state of the system. We investigated some variations of the state dependent prediction problem via the method of simulation. Our experimental results demonstrated that current state information can be useful in predicting the likelihood of a state in the near future, but its impact gradually decreases as the time difference between the present and the future grows larger. When extra knowledge of current patients' departure times is assumed, we showed that the ICU can be informed of the probability that the unit will be full at a certain future point in time by considering the number of current patients who will leave the ICU after that time, regardless of the number of total patients at the current moment.

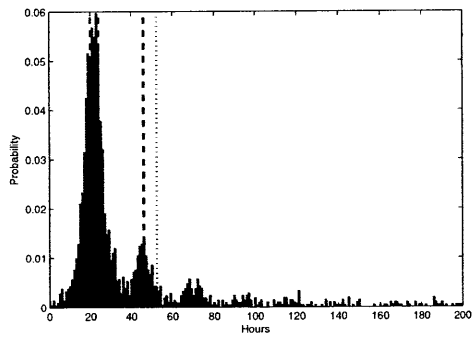
Several of our findings suggest further work. With respect to the current patients' LOS, it would be of interest to study the actual ability of the ICU staff to predict the remaining LOS of each current patient. The time window within which the ICU can precisely identify its patients' departures in reality can be used to design a more

realistic computational experiment for studying the state-dependent prediction problem. Furthermore, one could adopt the framework of state-dependent predictions to develop adaptive cap-based policies that adjust caps according to prediction results to avoid the future adverse events in the ICU. The typical policies would aim to prevent the chance of system overcrowding while being able to keep the system under stability and maintain a reasonable length of the mean waiting time of scheduled patients. Another interesting related direction is to formulate a continuous-time counterpart of the queueing model for the ICU at CHB. Not only would the model allow us to incorporate more complex policies and details into ICU systems, but it could enable the derivation of closed-form solutions, through which we can obtain immediate estimates for performance measures as well as a better insight into system behavior.

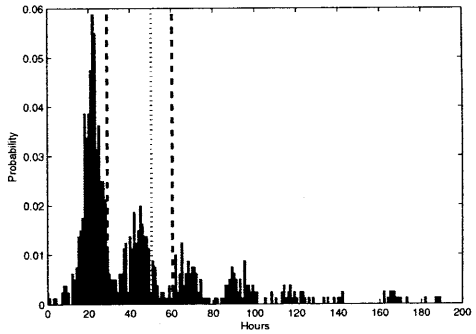
Appendix A

Additional Statistics

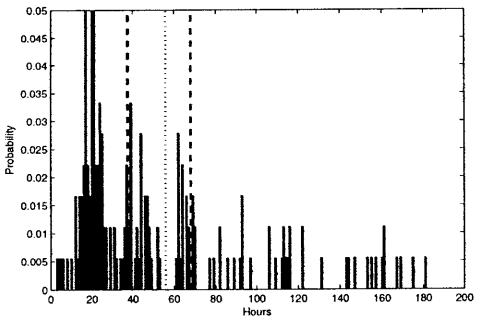
A.1 Distributions of the Length of Stay



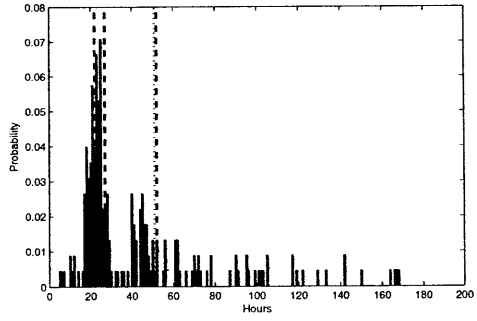
(a) ORL



(b) Plastics Surgery

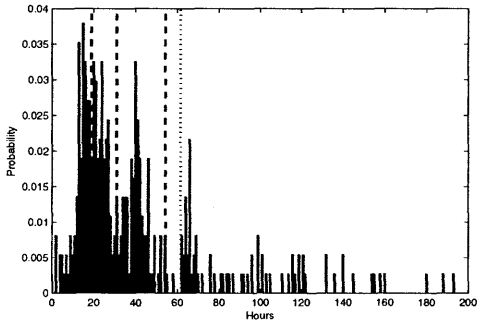


(c) Urology

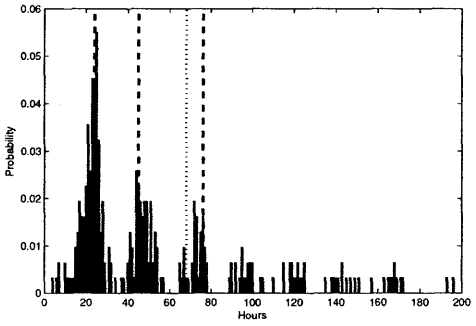


(d) OMFS

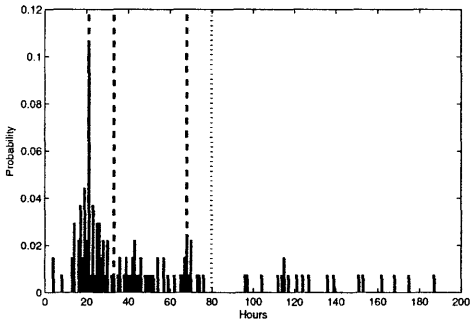
Figure A-1



(e) Trauma



(f) IntRadio



(g) Other surgery

Figure A-1: The distributions of the LOS by services from 1998-2008. The first three dashed lines indicate the first, second, and third quantiles respectively, while the dotted line locates the mean of the LOS. The horizontal axis is truncated to 200 hours.

A.2 Length of Stay by Types of Services from 1998-2008

Year	Mean LOS (hours)										
	Neurosurgical	ORL	Plastics	Urology	OMFS	Orthopedic	Trauma	IntRadio	General surgery	Other surgery	Medical
1998	74.68	36.81	97.71	-	-	72.06	-	-	92.90	72.50	106.66
1999	45.20	56.64	46.81	44.34	-	63.61	74.79	57.11	128.56	59.57	92.46
2000	58.33	44.20	44.47	44.24	-	54.40	46.52	59.39	109.79	42.82	121.85
2001	59.32	70.94	56.05	90.86	72.61	66.13	46.21	73.54	103.17	43.34	108.07
2002	39.04	64.68	48.19	43.69	68.24	88.82	61.81	80.47	105.12	42.54	97.10
2003	44.46	58.31	48.95	58.81	89.66	94.35	61.12	71.17	111.88	149.79	94.02
2004	52.01	46.75	46.21	34.79	36.47	81.17	74.73	74.83	101.39	25.64	109.73
2005	53.37	44.28	50.75	78.03	54.05	87.11	74.13	50.44	128.05	47.59	122.33
2006	52.88	60.44	53.77	66.31	41.06	105.09	70.25	68.26	166.48	125.02	119.77
2007	51.45	57.93	43.03	22.52	47.30	78.83	37.41	61.80	117.11	102.74	124.91
2008	50.97	42.74	57.79	61.59	37.13	74.73	60.99	79.46	147.73	56.73	120.86

Table A.1: Mean of the LOS by types of services from 1998-2008

Year	SD of the LOS (hours)										
	Neurosurgical	ORL	Plastics	Urology	OMFS	Orthopedic	Trauma	IntRadio	General surgery	Other surgery	Medical
1998	155.81	24.73	210.57	-	-	75.47	-	-	118.19	0	184.97
1999	74.21	108.50	41.26	33.49	-	118.95	125.80	39.49	220.20	57.65	136.11
2000	125.52	60.98	41.18	43.90	-	81.63	40.49	59.92	235.23	36.49	257.46
2001	121.52	118.22	45.84	104.02	39.95	107.15	32.65	102.07	157.39	48.80	201.81
2002	55.07	142.62	80.49	41.92	88.79	151.27	101.36	65.23	190.46	1.47	189.91
2003	57.45	101.60	33.24	42.36	120.16	135.25	66.32	50.28	205.82	395.56	173.42
2004	72.75	56.77	46.35	29.92	20.76	148.62	123.35	138.48	192.44	10.19	184.74
2005	88.93	62.45	52.75	59.61	79.54	130.99	148.44	40.77	272.24	75.44	230.24
2006	90.79	79.63	53.23	64.16	34.90	184.71	158.66	23.77	337.32	159.35	221.39
2007	86.16	103.31	41.33	14.50	50.44	106.52	36.86	58.93	212.69	315.97	233.24
2008	77.74	78.80	86.18	45.94	28.02	128.30	115.65	117.08	317.89	53.29	219.91

Table A.2: SD of the LOS by types of services from 1998-2008

Year	Total patients	Number of patients										
		Neurosurgical	ORL	Plastics	Urology	OMFS	Orthopedic	Trauma	IntRadio	General surgery	Other surgery	Medical
1998	721	33	20	19	0	0	15	0	0	31	1	602
1999	1566	201	117	104	22	0	133	39	15	224	14	697
2000	1578	230	143	90	25	0	148	36	39	238	8	621
2001	1587	226	126	61	26	7	115	44	27	230	4	721
2002	1697	211	146	66	17	39	128	36	33	279	2	740
2003	1832	266	193	97	25	12	157	30	37	233	11	771
2004	1661	248	208	70	12	28	135	42	38	175	5	700
2005	1550	259	187	69	12	42	133	35	29	204	8	572
2006	1611	265	193	66	16	28	129	51	9	220	13	621
2007	2102	309	282	67	12	26	160	28	37	317	36	828
2008	2099	297	384	83	11	38	191	25	41	348	23	658

Table A.3: Number of arrivals/admissions by types of services from 1998-2008. These numbers represent only patients that are used to analyze the statistics of the LOS.

Year	Percentage of patients (%)										
	Neurosurgical	ORL	Plastics	Urology	OMFS	Orthopedic	Trauma	IntRadio	General surgery	Other surgery	Medical
1998	4.58	2.78	2.64	0.00	0.00	2.08	0.00	0.00	4.31	0.14	83.61
1999	12.84	7.48	6.65	1.41	0.00	8.50	2.49	0.96	14.31	0.89	44.54
2000	14.58	9.07	5.71	1.59	0.00	9.38	2.28	2.47	15.09	0.51	39.38
2001	14.34	7.99	3.87	1.65	0.44	7.30	2.79	1.71	14.59	0.25	45.75
2002	12.47	8.63	3.90	1.00	2.30	7.57	2.13	1.95	16.49	0.12	43.74
2003	14.56	10.56	5.31	1.37	0.66	8.59	1.64	2.03	12.75	0.60	42.20
2004	14.98	12.57	4.23	0.73	1.69	8.16	2.54	2.30	10.57	0.30	42.30
2005	16.73	12.08	4.46	0.78	2.71	8.59	2.26	1.87	13.18	0.52	36.95
2006	16.47	12.00	4.10	0.99	1.74	8.02	3.17	0.56	13.67	0.81	38.60
2007	14.70	13.42	3.19	0.57	1.24	7.61	1.33	1.76	15.08	1.71	39.39
2008	14.15	18.29	3.95	0.52	1.81	9.10	1.19	1.95	16.58	1.10	31.35

Table A.4: Percentage of patients by types of services from 1998-2008. These numbers represent only patients that are used to analyze the statistics of the LOS.

Appendix B

Generating Random Variables from Empirical Data

In this appendix, we describe a method to generate samples from empirical data. The technique is used in drawing random values from any given set of data in the simulation model, including the length of stay (LOS) and the admission time of the day (W_{OR}) data.

Let U be a random variable that is uniformly distributed on the interval $[0, 1]$ and let F be the distribution function of a discrete random variable taking values $a_1, a_2, \ldots, a_n$.

Theorem 1. *The random variable X given by*

$$X = a_k \quad \textit{if} \quad F(a_{k-1}) < U \leq F(a_k)$$

has the cumulative probability function (CDF) F_X satisfying $F_X = F$.

Proof. For any $k = 1, 2, \ldots, n$, since

$$\Pr(F(a_{k-1}) < U \leq F(a_k)) = F(a_k) - F(a_{k-1}),$$

it follows that $\Pr(X = a_k) = F(a_k) - F(a_{k-1})$. This implies that $F_X = F$. □

The theorem indicates that the inverse-transform method can be used to generate a random variable from any given discrete distribution function F. In fact, a random variable can also be generated from any given continuous distribution function by means of the inverse transform (Grimmett and Stirzaker [15]).

A random *sample* x can be generated from a given F as follows. First, draw a random value u from U. Then, find a_k such that

$$F(a_{k-1}) < u \leq F(a_k),$$

and simply set $x = a_k$. Figure B-1 illustrates an example of drawing a random value using the inverse-transform method. See Banks et al. [3] for sampling methods for several types of distribution functions.

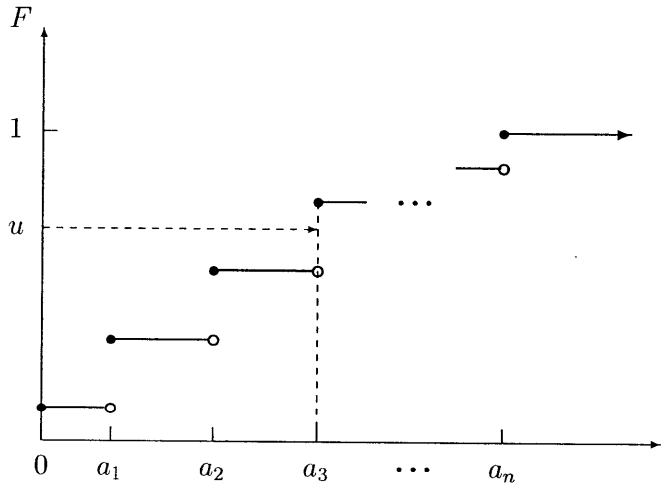


Figure B-1: Use of the inverse transform method for sampling from a discrete distribution F. Let u be a random sample drawn from U, which is uniformly distributed on $[0, 1]$. In this case, the sample value that corresponds to u is a_3.

Now, given a set of empirical data, we can construct the corresponding discrete distribution function based on the frequency with which each data point occurs. A random value can therefore be sampled from such a distribution by means of inversion as described above.

Bibliography

[1] M. Asaduzzaman and T. J. Chaussalet. Modelling and performance measure of a perinatal network centre in the united kingdom. In *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 506–511, 2005.

[2] M. Asaduzzaman, T. J. Chaussalet, and N. J. Robertson. A loss network model with overflow for capacity planning of a neonatal unit. *Ann Oper Res*, 178:67–76, 2010.

[3] J. Banks, J. S. Carson II, B. L. Nelson, and D. M. Nicol. *Discrete-Event System Simulation, 4th ed.* Prentice Hall, New Jersey, 2005.

[4] C. W. Chan, V. F. Farias, N. Bambos, and G. J. Escobar. Maximizing throughput of hospital intensive care units with patient readmissions. Submitted, available at `http://web.mit.edu/ vivekf/www/papers/icuV1.pdf`.

[5] T. J. Chaussalet, H. Xie, and P. Millard. A closed queueing approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*, 5:492–497, 2006.

[6] F. Dexter and R. Traub. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesth. Analg.*, 94:4933–4942, 2002.

[7] F. Dexter, A. Macario, R. Traub, M. Hopwood, and D. Lubarsky. An operating room scheduling strategy to maximize the use of operating room block time:

computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesth. Analg.*, 89:7–20, 1999.

[8] G. Dobson, H. H. Lee, and E. J. Pinker. Patient flow in an ICU. Technical Report FR 08-21, Univeristy of Rochester, 2008.

[9] R. B. Ferreira, F. C. Coelli, W. C. Pereira, and R. M. Almeida. Optimizing patient flow in a large hospital surgical centre by means of discrete-event computer simulation models. *Journal of Evaluation in Clinical Practice*, 14:1031–1037, 2008.

[10] R. G. Gallager. *Discrete Stochastic Processes.* Kluwer Academic Publishers, Norwell, Massachusetts, 1996.

[11] L. V. Green. How many hospital beds? *Inquiry*, 39(4):400–412, 2002.

[12] L. V. Green. *Patient Flow: Reducing Delay in Healthcare Delivery*, chapter Queueing analysis in Healthcare. Springer, New York, NY, 2006.

[13] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13:61–68, 2006.

[14] J. D. Griffiths and N. Price-Lloyd. A queueing model of activities in an intensive care unit. *IMA Journal of Management Mathematics*, 17:277–288, 2006.

[15] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes, 3rd ed.* Oxford University Press, Oxford, 2006.

[16] D. A. Gruenberg, W. Shelton, S. L. Rose, A. E. Rutter, S. Socaris, and G. McGee. Factors influencing length of stays in the intensive care unit. *American Journal of Critical Care*, 15:502–509, 2006.

[17] L. Jiang and R. E. Giachetti. A queueing network model to analyze the impact of parallelization of care on patient cycle time. *Health Care Management Science*, 11:248–261, 2008.

[18] S. C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research*, 115:36–46, 1999.

[19] S. C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, 18:427–443, 2000.

[20] L. Kleinrock. *Queueing Systems. Volume 1: Theory.* Wiley-Interscience, New York, NY, 1975.

[21] N. Koizumi, E. Kuno, and T. E. Smith. Modelling patient flows using a queueing network. *Health Care Management Science*, 8(1):49–60, 2005.

[22] A. Kolker. Process modeling of emergency department patient flow: effect of patient length of stay on ED diversion. *J. Med. Syst.*, 32:389–401, 2008.

[23] A. Kolker. Process modeling of ICU patient flow: effect of daily load leveling of elective surgeries on ICU diversion. *J. Med. Syst.*, 33:27–40, 2009.

[24] N. Kortbeek and N. M. van Dijk. On dimensioning intensive care units. *AENORM*, 57:22–26, 2007.

[25] J. F. Lawless. *Statistical Models and Methods for Lifetime Data, 2nd ed.* Wiley-Interscience, New Jersey, 2003.

[26] N. Litvak, M. van Rijsbergen, R. J. Boucherie, and M. van Houdenhoven. Managing the over flow of intensive care patients. *European Journal of Operational Research*, 185(3):998–1010, 2008.

[27] J. Lowery. Design of hospital admissions scheduling system using simulation. In J. Charness. and D. Morrice, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 1199–1204, 1996.

[28] A. Marazzi, F. Paccuad, C. Ruffieux, and C. Beguin. Fitting the distributions of length of stay by parametric models. *Medical Care*, 36(6):915–927, 1998.

[29] M. L. McManus, M. C. Long, A. Cooper, J. Mandell, D. M. Berwick, M Pagano, and E. Litvak. Variability in surgical caseload and access to intensive care services. *Anesthesiology*, 98:1491–1496, 2003.

[30] M. L. McManus, M. C. Long, A. Cooper, and E. Litvak. Queueing theory accurately models the need for critical care resources. *Anesthesiology*, 100:1271–1276, 2004.

[31] J. C. Ridge, S. K. Jones, M. S. Nielsen, and A. K. Shahani. Capacity planning for intensive care units. *European Journal of Operational Research*, 105:346–355, 1998.

[32] F. C. Ryckman, P. A. Yelton, A. M. Anneken, P. E. Kiessling, P. J. Schoettker, and U. R. Kotagal. Redesigning intensive care unit flow using variability management to improve access and safety. *The Joint Commission Journal on Quality and Patient Safety*, 35(11):535–543, 2009.

[33] K. Stricker, H. U. Rothen, and J. Takala. Resource use in the ICU: short- vs long-term patients. *Acta Anaesthesiol Scand.*, 47:508–515, 2003.

[34] P. M. Troy and L. Rosenberg. Using simulation to determine the need for ICU beds for surgery patients. *Surgery*, 146(4):608–620, 2009.

[35] J. B. Tucker, J. E. Barone, J. Cecere, R. G. Blabey, and C. K. Rha. Using queueing theory to determine operating room staffing needs. *J. Trauma*, 46: 71–79, 1999.

[36] D. C. Tyler, C. A. Pasquariello, and C. H. Chen. Determining optimum operating room utilization. *Anesth. Analg.*, 96:41114–41121, 2003.

[37] M. van Houdenhoven, J. M. van Oostrum, G. Wullink E. W. Hans, and G. Kazemier. Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. *Anesth. Analg.*, 105:3707–3714, 2007.

[38] N. Yankovic and L. V. Green. A queueing model for nurse staffing. Working Paper, Columbia University, Columbia Business School.